AUTODESK UNIVERSITY

IT21325

A Hardware Wonk's Guide to Specifying the Best 3D and BIM Workstations, 2016 Edition

Matthew Stachoni
BIM Specialist, Microsol Resources

Learning Objectives

- Understand the relative computing demands of Autodesk's leading BIM and 3D applications
- Learn the current state of the art and "sweet spots" in processors, memory, storage, and graphics
- Specify workstations for different classes of BIM and 3D usage profiles
- Understand how to shop for complete systems and individual components

Description

Working with today's leading Building Information Modeling (BIM) and 3D visualization tools presents a special challenge to your IT infrastructure. Wrestling with the computational demands of the Revit software platform, as well as BIM-related applications such as 3ds Max, Navisworks, Rhino, Lumion, and others, means that one needs the right knowledge to make sound investments in workstation hardware. This class gets inside the mind of a certified (or certifiable) hardware geek to understand the variables to consider when purchasing hardware to support the demands of these BIM and 3D applications.

Fully updated for 2016, this class will give you the scoop on the latest advancements in workstation processors, motherboards, memory, and graphics cards. This year we will pay special attention to the latest advancements in graphics technologies to meet the demands of high-end rendering, animation, and visualization across a wide variety of platforms

Your AU Expert

Matt Stachoni *has over 25 years of experience as a BIM, CAD, and IT manager for a variety of architectural and engineering firms, and has been using Autodesk software professionally since 1987. Matt is currently a BIM Specialist with Microsol Resources, an Autodesk Premier Partner in New York City, Philadelphia, and Boston. He provides training, BIM implementation, specialized consultation services, and technical support for all of Autodesk's AEC applications.*

Previously, Matt was the BIM and IT Manager for Erdy McHenry Architecture LLC and a BIM specialist for CADapult Ltd. (now Applied Software), where he provided on-site construction BIM modeling and coordination services for construction managers, HVAC, and MEP trade contractors for a number of projects. He is a contributing writer for AUGIWorld Magazine. This is his 18th consecutive year attending Autodesk University and his 14th year as a speaker.

email: *matt@stachoni.com*
mstachoni@microsolresources.com

Twitter: @MattStachoni

I. Introduction

"Wow, this workstation is just way too fast for me."
– No one. Ever.

Specifying new BIM / 3D workstations, particularly ones tuned for Autodesk's 3D and BIM applications, can be a daunting task with all of the choices you have. You can spend quite a bit of research wading through online reviews, forums, and talking with salespeople who don't understand what you do on a daily basis. Moreover, recent advancements in both hardware and software often challenge preconceptions of what is important.

Computing hardware had long ago met the relatively low demands of 2D CAD, but data-rich 3D BIM and visualization processes will tax any workstation to some extent. Many of the old CAD rules no longer apply; you are not working with small project files, as individual project assets can exceed a Gigabyte as the BIM data grows and modeling gets more complex. The number of polygons in your 3D views in even modest models can be huge. Additionally, Autodesk's high-powered BIM and 3D applications do not exactly fire up on a dime.

Today there exists a wide variety of tools to showcase BIM projects, so users who specialize in visualization will naturally demand the most powerful workstations you can find. However, the software barrier to entry for high end visualization results is dropping dramatically, as we are seeing modern applications that are both easy to learn and create incredible photorealistic images.

The capability and complexity of the tools in Autodesk's various Suites and Collections improves with each release, and those capabilities can take their toll on your hardware. Iterating through adaptive components in Revit, or using the advanced rendering technologies such as the Iray rendering engine in 3ds Max will tax your workstation's subsystems differently. Knowing how to best match your software challenges in hardware is a key aspect of this class.

Taken together, this class is designed to arm you with the knowledge you need to make sound purchasing decisions today, and to plan for what is coming down the road in 2017.

What This Class Will Answer

This class will concentrate on specifying new workstations for BIM applications in the Autodesk® Building Design Suite and Architecture, Engineering, and Construction Industry Collection. Most notably we focus on Revit® and 3ds Max® performance, as well as third-party AEC tools such as Lumion 3D, Rhino, Vray, Arnold, Unreal Engine 4, and others. Moreover, the concepts provided herein readily stretch across all disciplines and areas of work, including mechanical engineering, manufacturing, factory design, and fabrication.

Ideally, we want to look at these fundamental concerns:

- Identify typical user profiles that have specific hardware demands.
- Identify the latest software trends in the AEC industry and how they affect hardware decisions.
- Identify the latest computing hardware trends, both in technology and cost.

This leads to a more specific series of questions regarding system component specifications:

- What do the latest round of CPUs provide in terms of value and performance?
- How much system RAM is appropriate for my applications? Where does it make a difference?
- What is the difference between a workstation graphics card and a "gaming" graphics card?
- Are solid state drives (SSDs) worth the extra cost? What size is currently appropriate?
- Do I build my own machine or do I buy a complete system from a vendor?
- Given the current state of the art, should I consider upgrading or replacing my hardware outright?

Accordingly, this class covers the five critical subsystems and review specific components found in every workstation: Central Processing Units (CPUs); chipsets and motherboard features; system memory (RAM); graphics processors (GPUs); and storage. In addition we will look at what is new in laptop and mobile platforms, peripherals, operating systems, and include a comprehensive 2016-2017 Buying Guide.

Disclaimer

In this class I will often make references and tacit recommendations for specific system components. These are purely my opinion, stemming largely from extensive personal experience and research in building systems for myself, my customers, and my company. Use this handout as a source of technical information and a buying guide, but remember that you are spending your own money (or the money of someone you work for). Thus, the onus is on you to do your own research when compiling your specifications and systems. I have no vested interest in any component manufacturer and make no endorsements of any specific product mentioned in this document.

Identifying Your User Requirements

The first thing to understand is that one hardware specification does not fit all user needs. You must understand your users' specific computing requirements. In general I believe we can classify users into one of three use-case scenarios and outfit them with a particular workstation profile.

1. **The Grunts:** These folks use Revit day in and day out, and rarely step outside of that to use more sophisticated software. They are typically tasked with the mundane jobs of project design, documentation, and project management, but do not regularly create complex, high end renderings or extended animations. Revit is clearly at the top of the process-consumption food chain, and nothing else they do taxes their system more than that. However, many Grunts will evolve over time into more complex workloads, so their workstations need to handle at least some higher-order functionality without choking.

2. **The BIM Champs:** These are your BIM managers and advanced users who not only use Revit all day for production support, but delve into the nooks and crannies of the program to help turn the design concepts into modeled reality. They not only develop project content, but create Dynamo scripts, manage models from a variety of sources, update and fix problems, and so on. BIM Champs may also regularly interoperate with additional 3D modeling software such as 3ds Max, Rhino, Lumion, and SketchUp, and pull light to medium duty creating visualizations. As such their hardware needs are greater than those of the Grunt, although perhaps in targeted areas.

3. **The Viz Wizards:** These are your 3D and visualization gurus who may spend as much time in visualization applications as they do in Revit. They routinely need to push models in to and out of 3ds Max, Rhino, Maya, InfraWorks 360, SketchUp, and others. They run graphics applications such as Adobe's Photoshop, Illustrator, and others - often concurrently with Revit and 3ds Max. They may extensively use real-time ray tracing found in Unreal Engine 4 and Lumion. These users specialize in photorealistic renderings and animations, and develop your company's hero imagery. The Viz Wiz will absolutely use as much horsepower as you can throw at them.

Ideally, each one of these kinds of users would be assigned a specific kind of workstation that is fully optimized for their needs. Given that you may find it best to buy systems in bulk, you may be tempted to specify a single workstation configuration for everyone without consideration to specific user workloads. I believe this is a mistake, as one size does not fit all. On the other hand, large disparities between systems can be an IT headache to maintain. Our goal is to establish workstation configurations that target these three specific user requirement profiles.

II. Industry Pressures and Key Trends

In building out any modern workstation or IT system, we need to first recognize the size of the production problems we are working with, and understand what workstation subsystems are challenged by a particular task. Before we delve too deeply into the specifics of hardware components, let's review some key hardware industry trends which shape today's state of the art and drive the future of computing:

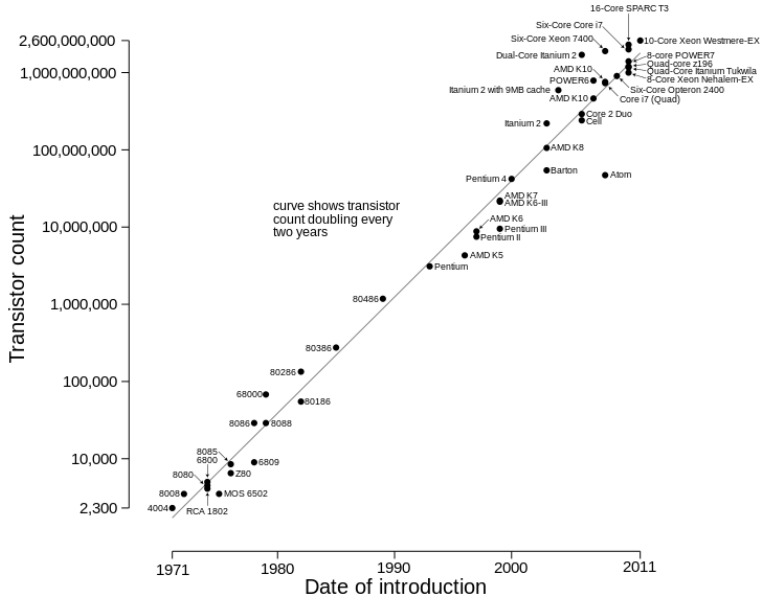
- Maximizing Performance per Watt (PPW);
- The slowdown of yearly CPU performance advancement and the potential end of Moore's Law;
- Realizing the potential that parallelism, multithreading, and multiprocessing bring to the game;
- Understanding the impact of PC gaming and GPU-accelerated computing for general design;
- Increased adoption of virtualization and cloud computing;
- Tight price differentials between computer components

Taken together these technologies allow us to scale workloads up, down, and out.

Maximizing Performance per Watt and Moore's Law

Every year Intel, Nvidia, and AMD release new iterations of their hardware, and every year their products get faster, smaller, and cooler. Sometimes by a little, sometimes by a lot. Today, a key design criteria used in today's microprocessor fabrication process is to maximize energy efficiency, measured in Performance per Watt, or PPW.

For years, the rate of improvement in integrated circuit design has been predicted quite accurately by Gordon E. Moore, a co-founder of Intel. ***Moore's Law,*** first coined in his 1956 paper "*Cramming more components onto integrated circuits*[1]," is the observation that, over the history of computing hardware, the number of transistors in an integrated circuit has roughly doubled approximately every two years[2].



Semiconductor manufacturing processes

10 µm – 1971
6 µm – 1974
3 µm – 1977
1.5 µm – 1982
1 µm – 1985
800 nm – 1989
600 nm – 1994
350 nm – 1995
250 nm – 1997
180 nm – 1999
130 nm – 2001
90 nm – 2004
65 nm – 2006
45 nm – 2008
32 nm – 2010
22 nm – 2012
14 nm – 2014
10 nm – 2017
7 nm – ~2019
5 nm – ~2021

Figure 1 - Transistor Count and Moore's Law, from 1971 to 2011. Note the logarithmic vertical scale.

[1] https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/view?usp=sharing
[2] Moore's Law, https://en.wikipedia.org/wiki/Moore%27s_law

How Transistors Work

A transistor is, at its heart, a relatively simple electrically driven switch that controls signaling current between two points. When the switch is open, no current flows and the signal has a value of 0. When the switch is closed, the current flows and you get a value of 1. We combine transistors together into larger circuits that can perform logical operations. Thus, the number of transistors on a processor directly determines what that chip can do, so cramming more of them in a certain amount of space is a critical path to performance improvement.

The most common transistor design is called a ***metal-oxide-semiconductor field-effect transistor***, or MOSFET, which is a building block of today's integrated circuits. Fundamentally, a MOSFET transistor has four parts: a source, a drain, a channel that connects the two, and a gate on top to control the channel. When the control gate has a positive voltage applied to it, it generates an electrical field that attracts negatively charged electrons in the channel underneath the gate, which then becomes a conductor between the source and drain. The switch is turned on.

Making transistors smaller is primarily accomplished by shrinking the space between the source and drain. This space is determined by the semiconductor ***technology node*** using a particular lithography fabrication process. A node/process is measured in nanometers (nm), or millionths of a millimeter.

Moore's law, being an exponential function, means the rate of change is always increasing. This has largely been true until just recently. Every two to four years a new, smaller technology node makes its debut and the fabrication process has shrunk from 10,000 nm (10 microns) wide in 1971 to only 14 nm wide today. To give a sense of scale, a single human hair is about 100,000 nm (100 microns) wide. Moving from 10,000nm to only 14nm is equivalent to shrinking a 5'-6" tall person down to the size of a grain of rice.

Accordingly, transistor count has gone up from 2,300 transistors to somewhere between 1.35 - 2.6 billion transistors in today's CPU models. Think about this: Boston Symphony Hall holds about 2,370 people (during Pops season). The population of China is about 1.357 billion people. Now squeeze the entire population of China into Boston Symphony Hall. That's Moore's Law for the past 45 years.

As a result of a smaller fabrication process, integrated circuits use less energy and produces less heat, which also allow for more densely packed transistors on a chip. In the late 1990s into the 2000s the trend was to increase on-die transistor counts and die sizes, but with the fabrication process still in the 60nm to 90nm range, CPUs simply got a lot larger. Energy consumption and heat dissipation became serious engineering challenges, and led to a new market of exotic cooling components such as large fans, CPU coolers with heat pipes, closed-loop water cooling solutions with pumps, reservoirs and radiators, and even submerging the entire PC in a vat of mineral oil. Clearly, the future of CPU microarchitectures depended on shrinking the fabrication process for as long as technically possible.

Today's 14nm processors and 14-16nm GPUs are not only physically smaller, but also have advanced internal power management optimizations that reduce power (and thus heat) when it is not required. Increasing PPW allows higher performance to be stuffed into smaller packages and platforms, which opened the floodgates to the vast development of mobile technologies that we all take for granted.

This had two side effects. First, the development of more powerful, smaller, cooler running, and largely silent CPUs and GPUs allows you to stuff more of them in a single workstation without it cooking itself. At the same time, CPU clock speeds have been able to rise from about 2.4GHz to 4GHz and beyond.

Secondly, complex BIM applications can now extend from the desktop to mobile platforms, such as actively modeling in 3D using a small laptop during design meetings, running clash detections at the construction site using tablets, or using drone-mounted cameras to capture HD imagery.

Quantum Tunneling and the Impending End of Moore's Law

While breakthroughs in MOSFET technology have enabled us to get down to a 14nm process, we are starting to see the end of Moore's law on the horizon. The space between the source and drain at 14nm is only about 70 silicon atoms wide. At smaller scales, the ability to control current flow across a transistor without leakage becomes a significant problem.

By 2026 we expect to get down to a 5nm process, which is only about 25 atoms wide. This 5nm node is often assumed to be the practical end of Moore's law, as transistors smaller than 7nm will experience an increase in something called "quantum tunneling" which impacts transistor function. Quantum tunneling is the weird effect that happens when the process becomes so small that the probability of electrons simply passing through the logic gate barrier increases and becomes a source of leakage, keeping the switch from doing its job and thus limiting the size where the information being passed is still completely reliable. To fix this scientists have come up with 3D gate designs which are tall enough to minimize the probability of quantum tunneling, but the pace to move downward is slowing. To paraphrase Intel Fellow Mark Bohr, we are simply quickly running out of atoms to play with.[3]

In the end, however, the future of microprocessor design will need to rely much less on shrinking the process, but through clever and innovative rethinking of microarchitectures and superscalar system design. But these kinds of improvements will likely be much less dramatic that what we have traditionally experienced over recent years. In fact, our discussion on the latest Intel CPUs reflect exactly this trend.

Parallel Processing, Multiprocessing, and Multithreading

It has long been known that key problems associated with BIM and 3D visualization, such as energy modeling, photorealistic imagery, and engineering simulations are simply too big for a single processor to handle efficiently. Many of these problems are highly parallel in nature, where large tasks can often be neatly broken down into smaller ones that don't rely on each other to finish before the next one can be worked on. This lead to the development of operating systems that support multiple CPUs.

First, some terminology on CPUs and cores. According to Microsoft, "systems with more than one physical processor or systems with physical processors that have multiple cores provide the operating system with multiple logical processors. A *logical processor* is one logical computing engine from the perspective of the operating system, application or driver. A core is one processor unit, which can consist of one or more logical processors. A physical processor can consist of one or more cores. A physical processor is the same as a processor package, a socket, or a CPU."[4]

In other words, an operating system such as Windows 10 will see a single physical CPU that has four cores as four separate logical processors, each of which can have threads of operation scheduled and assigned. The 64-bit versions of Windows 7 and later support more than 64 logical processors on a single computer. This functionality is not available in 32-bit versions of Windows.

[3] http://www.theregister.co.uk/2013/08/27/moores_law_will_be_repealed_due_to_economics_not_physics/
[4] https://msdn.microsoft.com/en-us/library/windows/desktop/dd405503(v=vs.85).aspx

All modern processors and operating systems fully support both **multiprocessing -** the ability to push separate processes to multiple CPU cores in a system - and **multithreading,** the ability to execute separate threads of a single process across multiple processors. Processor technology has evolved to meet this demand, first by allowing multiple physical CPUs on a motherboard, then by introducing more efficient multi-core designs in a single CPU package. The more cores your machine has, the snappier your overall system response is and the faster any compute-intensive task such as rendering will complete.

These kinds of non-sequential workloads can be distributed to multiple processor cores on a CPU, multiple physical CPUs in a single PC, or even out to multiple physical computers that will chew on that particular problem and return results that can be aggregated later. Over time we've all made the mass migration to multi-core computing even if we aren't aware of it, even down to our tablets and phones.

In particular, 3D photorealistic rendering lends itself very well to parallel processing. The ray tracing pipeline used in today's rendering engines involves sending out rays from various sources (lights and cameras), accurately bouncing them off of or passing through objects they encounter in the scene, changing the data "payload" in each ray as it picks up physical properties from the object(s) it interacts with, and finally returning a color pixel value to the screen. This process is computationally expensive as it has to be physically accurate, and can simulate a wide variety of visual effects, such as reflections, refraction of light through various materials, shadows, caustics, blooms, and so on.

You can see this parallel processing in action when you render a scene using the mental ray rendering engine. mental ray renders scenes in separate tiles called *buckets*. Each processor core in your CPU is assigned a bucket and renders it before moving to the next one. The number of buckets you see corresponds to the number of cores available. The more cores, the more buckets, and the faster the rendering.

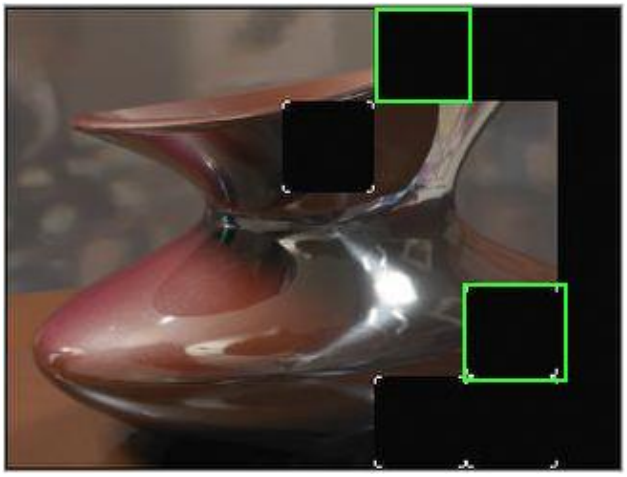


Image from "Mastering mental ray" by Jennifer O'Connor

Autodesk recognized the benefits of parallelization and provides the Backburner distributed rendering software with 3ds Max. You can create your own rendering farm where you send a rendering job out to multiple computers on your local area network, each of which would render a little bit of the whole, send their finished portion back, which then gets assembled back into a single image or animation. With enough machines, what would take a single PC hours can be created in a fraction of the time.

Just running an operating system and multiple concurrent applications is, in many ways, a parallel problem as well. Even without running any applications, a modern OS has many background processes running at the same time, such as the security subsystem, anti-virus protection, network connectivity, disk I/O, and the list goes on. Each of your applications may run one or more separate processes as well, and processes themselves can spin off separate threads of execution. For example, Revit's rendering process is separate from the host Revit.exe process. In AutoCAD, the VisualLISP subsystem runs in its own separate thread.

While today you can maximize efficiency for highly parallel CPU workloads by outfitting a workstation with multiple physical CPUs, each with multiple cores, this is significantly expensive and a case of diminishing returns. Other advancements may point to other directions instead of trying to pile on CPU cores.

The Road to GPU Accelerated Computing and the Impact of Gaming

Recognizing the parallel nature of many graphics tasks, graphic processor unit (GPU) designers at AMD and Nvidia have created microarchitectures that are massively multiprocessing in nature and are fully programmable to boot. Given the right combination of software and hardware, we can now offload compute-intensive parallelized portions of a problem to the graphics card and free up the CPU to run other code. In fact these new GPU-compute tasks do not have to be graphics related, but could model weather patterns, run acoustical analysis, perform protein folding, and work on other complex problems.

Fundamentally, CPUs and GPUs process tasks differently, and in many ways the GPU represents the future of parallel processing. GPUs are specialized for compute-intensive, highly parallel computation - exactly what graphics rendering is about - and are therefore designed such that more transistors are devoted to raw data processing rather than data caching and flow control.

A CPU consists of a few – from 2 to 8 in most systems - relatively large cores which are optimized for sequential, serialized processing, executing a single thread at a very fast rate, between 3 and 4GHz. Conversely, today's GPU has a massively parallel architecture consisting of thousands of much smaller, highly efficient cores designed to execute many concurrent threads more slowly – between 1 and 2 GHz.

The GPU's physical chip is also larger. With thousands of smaller cores, a GPU can have 3 to 4 times as many transistors on the die than a CPU. Indeed, it is by increasing the PPW that the GPU can cram so many cores into a single die.

Real Time Rendering in Gaming

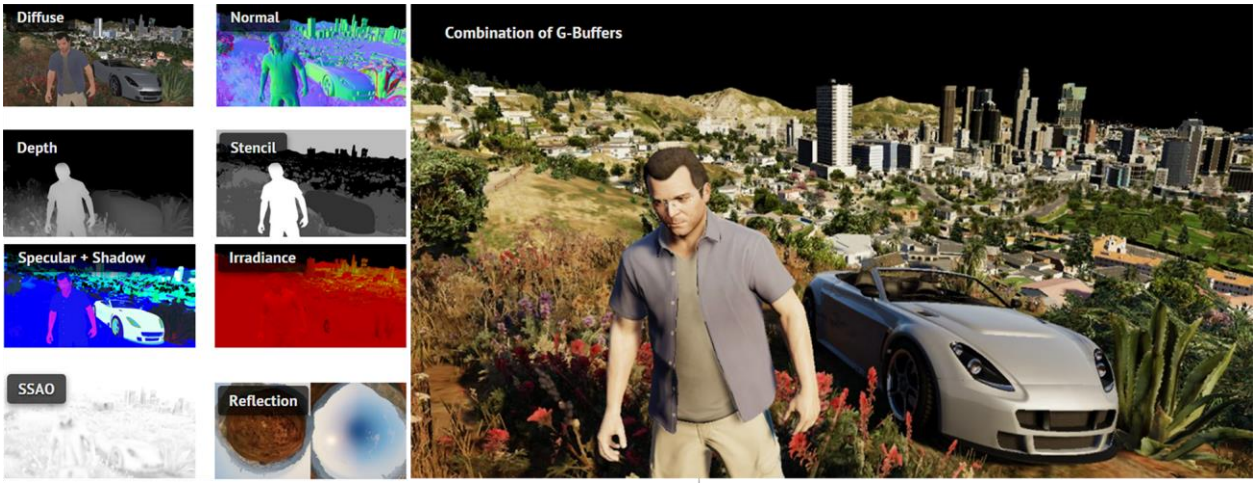
Back in olden times traditional GPUs used a fixed-function pipeline, and thus had a much more limited scope of work they could perform. They did not really think at all, but simply mapped function calls from the application through the driver to dedicated logic in the GPU that was designed to support them in a hard-coded fashion. This led to all sorts of video driver-related issues and false optimizations.

Today's graphics data pipeline is much more complex and intelligent. It is composed of a series of steps used to create a 2D raster representation from a 3D scene in real time. The GPU is fed 3D geometric primitive, lighting, texture map, and instructional data from the application. It then works to transform, subdivide, and triangulate the geometry; illuminate the scene; rasterize the vector information to pixels; shade those pixels; assemble the 2D raster image in the frame buffer; and output it to the monitor.

In games, the GPU needs to do this as many times a second as possible to maintain smoothness of play. For example, a detailed dissection of a rendered frame from *Grand Theft Auto V*[5] reveals a highly complex rendering pipeline. The 3D meshes that make up the scene are culled and drawn in lower and higher levels of detail depending on their distance from the camera. Even the lights that make up an entire city nighttime scene are individually modeled - that's tens of thousands of polygons being pushed to the GPU.

The rendering pipeline then performs a large array of multiple passes, rendering out many High Dynamic Range (HDR) buffers. These are screen-sized bitmaps of various types, such as diffuse, specular, normal, irradiance, alpha, shadow, reflection, etc. Along the way it applies effects for water surfaces, subsurface scattering, atmosphere, sun and sky, and transparencies. Then it applies tone mapping, i.e. photographic exposure, which converts the HDR information to a Low Dynamic Range (LDR) space. The scene is then anti-aliased to smooth out jagged edges of the meshes, a lens distortion is applied to make things more film-like, and the user interface (e.g. health, status, the mini-map of the city) is drawn on top of the scene. Then post effects such as lens flares, light streaks, anamorphic lenses, heat haze, and depth of field to blur out things that are not in focus are applied.

[5] http://www.adriancourreges.com/blog/2015/11/02/gta-v-graphics-study/



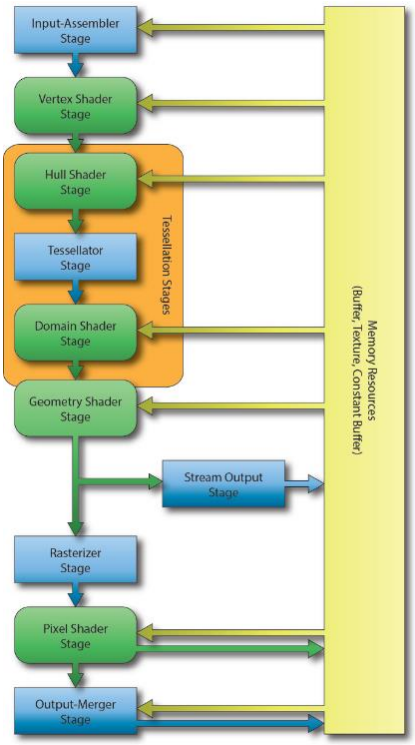
A game like GTA V needs to do all of this about 50 to 60 times a second to make the game playable. But how can all of these very highly complex steps be performed at such a high rate?

Shaders

Today's graphics pipelines are manipulated through small programs called ***Shaders***, which work on scene data to make complex effects happen in real time. Both OpenGL and Direct3D (part of the DirectX multimedia API for Windows) are 3D graphics APIs that went from the old-timey fixed-function hard-coded model to supporting the newer programmable shader-based model (in OpenGL 2.0 and DirectX 8.0).

Shaders work on a specific aspect of a graphical object and pass it on to the next step in the pipeline. For example, a Vertex Shader processes vertices, performing transformation, skinning, and lighting operations. It takes a single vertex as an input and produces a single modified vertex as the output. Geometry shaders process entire primitives consisting of multiple vertices, edges, polygons. Tessellation shaders subdivide simpler meshes into finer meshes allowing for level of detail scaling. Pixel shaders compute color and other attributes, such as bump mapping, shadows, specular highlights, and so on.

Shaders are written to apply transformations to a large set of elements at a time, which is very well suited to parallel processing. This dovetails with newer GPUs with many cores to handle these massively parallel tasks, and modern GPUs have multiple shader pipelines to facilitate high computational throughout. The DirectX API, released with each version of Windows, regularly defines new shader models which increase programming model flexibilities and capabilities.



The graphics pipeline for Microsoft DirectX 11

Modernizing Traditional Professional Renderers

Two of the primary 3D rendering engines in Autodesk's AEC collection of applications are Nvidia's mental ray and the new Autodesk Raytracer. With the recent acquisition of Solid Angle, 3ds Max and Maya now have the Arnold rendering engine as well, which may make it into Revit and other applications in the future. All support real-world materials and photometric lights for producing photorealistic images.

However, mental ray is owned and licensed by Nvidia, to which Autodesk pays a licensing fee with each application it ships with. Autodesk simply takes the core mental ray code and retrofits a User Interface around it for Revit, 3ds Max, etc.

Additionally, mental ray is almost 30 years old whereas the Autodesk Raytracer and, to a smaller extent, Arnold, are brand new. Both ART and Arnold are physically based renderers, whereas mental ray uses caching algorithms such as Global Illumination and Final Gather to simulate the physical world. As such both ART and Arnold ideal for interactive rendering via ActiveShade in 3ds Max.

For end users the primary difference between ART/Arnold and mr is in simplicity and speed, where these newer engines can produce images much faster, more efficiently, and with far less tweaking than mental ray. ART and Arnold also produce images that are arguably of better rendering quality[6]. Autodesk Raytracer is currently in use in AutoCAD, Revit, 3ds Max, Navisworks, and Showcase. Arnold ships with Maya and Arnold 0.5 (also called MAXtoA) is available as an preview release add-in for 3ds Max 2017.

CPU vs. GPU Rendering with Iray

However, neither mental ray, ART, Arnold, or other popular 3rd party renderers like V-Ray Advanced use the computational power of the GPU to accelerate rendering tasks. Rendering with these engines is almost entirely a CPU-bound process, so a 3D artist workstation would need to be outfitted with multiple (and expensive) physical multi-core CPUs. As mentioned previously, you can significantly lower render times in 3ds Max by throwing more PCs at the problem via setting up a render farm using the included Backburner software. However, each node on the farm needs to be pretty well equipped and Backburner's reliability through a heavy rendering session has always been shaky, to say the least. That has a huge impact on how you can easily manage rendering workloads and deadlines.

Designed for rasterizing many frames of simplified geometry to the screen per second, GPUs were not meant for performing ray-tracing calculations. This is rapidly changing as most of a GPU's hardware is now devoted to 32-bit floating point shader processors. Nvidia exploited this in 2007 with an entirely new GPU computing environment called **CUDA (Compute Unified Device Architecture),** which is a parallel computing platform and programming model established to provide direct access to the massive number of parallel computational elements in their CUDA GPUs. Non-CUDA platforms (that is to say, AMD graphics cards) can use the Open Computing Language (OpenCL) framework, which allows for programs to execute code across heterogeneous platforms – CPUs, GPUs, and others.

Using CUDA / OpenCL platforms, we have the ability to perform non-graphical, general-purpose computing on the GPU, often referred to as *GPGPU*, as well as accelerating graphics tasks such as calculating game physics.

One of the most compelling areas GPU Compute can directly affect Autodesk applications is with the **Nvidia Iray** rendering engine. Included with 3ds Max, Nvidia's Iray renderer fully uses the power of a CUDA-enabled (read: Nvidia) GPU to produce stunningly photorealistic imagery. We'll discuss this in more depth in the section on graphics. Given the nature of parallelism, I would not be surprised to see GPU compute technologies to be exploited for other uses across all future BIM applications.

[6] http://www.aecbytes.com/tipsandtricks/2015/issue72-revit.html

Using Gaming Engines for Architectural Visualization

Another tack is to exploit technology we have now. We have advanced shaders and relatively cheap GPU hardware that harnesses them, creating beautiful imagery in real time. So instead of using them to blow up demons on Mars or check some fool on the ice, why not apply them to the task of design visualization?

The advancements made in today's game engines is quickly competing with, and sometimes surpassing, what dedicated rendering engines like mental ray, v-ray and others can create. A game engine is a complete editing environment for working with 3D assets. You typically import model geometry from 3ds Max or Maya, then develop more lifelike materials, add photometric lighting, animations, and write custom programming code to react to gameplay events. Instead of the same old highly post-processed imagery or "sitting in a shopping cart being wheeled around the site" type animations, the result is a free running "game" that renders in real time, allowing you and your clients to explore and interact with. While 3D immersive games have been around for ages, the difference is that now the overall image quality in these new game engines is incredibly high and certainly good enough for design visualization.

For example, you may be familiar with Lumion, which is a very popular real-time architectural visualization application. Lumion is powered by the Quest3D 3D engine, which Act-3D developed long ago (before most gaming engines were commercially available) as a general 3D authoring tool, on top of which is lots of work with shaders and other optimizations, and easy UI, and lots of prebuilt content.

Currently the most well-known gaming engines available are Unreal Engine 4 and Unity 5, which are quickly becoming co-opted by the building design community. What's great about both is their cost to the design firm - they're **free**. Both Unreal and Unity charge game publishers a percentage of their revenue, but for design visualizations, there is no charge. The user community is growing every day, and add-ons, materials, models, and environments are available that you can purchase and drop into your project.

This is not a photograph. It's Unreal Engine 4. And it's real time. And it's free. Image courtesy ue4arch.com.

In 2014, Autodesk acquired the Bitsquid game engine which has evolved into their new Stingray[7] game engine, which is also cheap, as it is included with Maya LT can be subscribed to for a low $30/month fee.

Editing architectural visualizations using Autodesk Stingray

Virtual Machines, Virtual Desktop Infrastructure, and Cloud Computing

One of the more compelling side-effects of cheap, fast processing is the (re)rise of virtual computing. You may be familiar with Virtual Machine (VM) software such as Oracle's VirtualBox, Parallels Desktop, VMware Workstation, and Microsoft Hyper-V. These allow you to host a complete desktop OS on your PC, allowing you to run two operating systems at one time. This "computer within a computer" model emulates (or *abstracts*) physical hardware into a standard set of virtual hardware, which consists of virtual disks, virtual CPUs, virtual memory, virtual display, virtual USB ports, and so on.

VMs are in use in most businesses today in some fashion, typically for servers. Companies may host multiple server VMs on a single high-end box, which allows them to deploy fewer physical boxes to host file storage servers, Microsoft Exchange servers, SQL database servers, application servers, web servers, and others. Many firms put Revit Server on its own VM.

This is valuable because many server services don't require a lot of horsepower, but you don't usually want to combine application servers on one physical box under a single OS. You don't want your file server also hosting Exchange, for example, for many reasons; the primary one being that if one goes down it takes the other out with it. Putting all your eggs in one basket usually leaves you with scrambled eggs when you hit a bump in the road.

VMs also allows IT a lot of flexibility in how these servers are apportioned across available hardware and allows for better serviceability. Seen from the outside, VMs are just single files that contain the OS, files, and applications. As such a VM can be shut down independently of the host box or other VMs, moved to another machine, and fired up within minutes. You cannot do this with Microsoft Exchange or your company accounting system installed on a normal server.

[7] http://www.engadget.com/2015/08/03/autodesk-stingray/

Virtual Desktop Infrastructure (VDI) technology upscales this to serve the enterprise as a whole. Simply put, VDI is a data center technology that supplies hosted desktop images to remote users. Remote desktop virtualization is becoming very popular and is frequently used in the following scenarios:

- In environments with high availability requirements and where technical support is not readily available;
- In environments where high network latency degrades the performance of client/server applications;
- In environments where remote access and data security needs create conflicting requirements that can only be addressed by retaining all application data within the data center - with only display, keyboard, and mouse information communicated with the remote client.

All of these environments are seen in the AEC design space. Design firms often open smaller remote offices to serve local projects or expand their operations into new markets. These offices are often too small for dedicated on-premises technical support, and may not want the care and feeding overhead of an IT closet full of servers and the data therein. These remote offices can use VDI technology to access desktops hosted at the central office, where all data is kept, backed up, and readily accessible across the company.

Because all data and desktop processing is on the server, it reaps certain performance benefits as well. Users routinely hit performance bottlenecks as large central files residing on servers are delivered to the local machine. Concurrent synchronizations often compete for network resources, slowing down all LAN traffic throughout the day. VDI keeps all BIM data on the server all the time, providing local access speeds.

Furthermore, IT departments can more easily service the back-end hardware, as they can bounce user desktops from server to server – even when the session is running, invisible to the user. They provide well over 99% uptime, and can instantly deploy new applications and updates across the entire enterprise easily. There is a huge time sink in managing Autodesk software across many physical desktops, particularly with the large number of applications in today's Autodesk Suites and Collections as well as their continuous upkeep with patches, service packs, and updates.

The primary impediment to deploying Virtual Desktops of high-end applications like Revit, Navisworks, and 3ds Max has been in how to share 3D hardware-accelerated graphics between virtual machines. Older VMs could not provide the kind of dedicated virtual graphics capabilities required by these applications to run well. New technologies have emerged, such as Nvidia's GRID technology for sharing virtual GPUs across multiple virtual desktop and application instances.[8]

That said, it is important to understand that even the best VDI solutions will typically not perform as well as standalone, dedicated workstations under heavy usage. The hardware sits as a virtualized layer between the OS and the physical host, to allow the host hardware to be shared between several (sometimes many) virtual desktops. Because you are slicing up a single host workstation into smaller ones, you almost certainly have to shortchange the Virtual Desktops in some manner, such as reducing the number of virtual CPU cores, virtual RAM, virtual video RAM, virtual disk space, and so on. Moreover, to provide enough virtual processing power, VDI servers often use exotic, multicore CPUs with perhaps 10 to 20 cores, which run much slower than 4 to 8 core high end desktop CPUs. You cannot ever perform faster than the native hardware that underpins the VM, so single-threaded application performance often suffers at these slower speeds.

However, overall operational benefits noted above may simply trump local performance considerations. As with specifying single-user workstation hardware – the primary focus of this class - correctly sizing VDI servers and Virtual Desktops according to user requirements is key to ensuring application performance.

[8] http://www.nvidia.com/object/grid-technology.html

The Cloud Effect

VDI technologies have migrated from you hosting multiple virtual machines on your server(s) to someone else doing it for you, me, and everyone else. Data centers are everywhere and accessible any time. That is why they call it the Cloud, and no information technology discussion today would be complete with some reference to cloud computing.

By now, it's taken for granted that processing speed increases over time and price per process drops. This economy of scale is coupled with the ubiquitous adoption of very fast Internet access at almost every level. The mixing of cheap and fast computing performance with ubiquitous broadband networking has resulted in easy access to remote processing horsepower. Just as the cost of 1GB of disk storage has plummeted from $1,000 to just a few pennies, the same thing is happening to CPU cycles as they become widely available on demand.

This has manifested itself in the emerging benefit of widely distributed cloud-based computing services. The cloud is quickly migrating from the low hanging fruit of simple storage-anywhere-anytime mechanism (e.g., Dropbox, Box.net, even A360), to providing users with massive remote access capabilities to fast machines. This will soon become on-demand, essentially limitless, and very cheap computing horsepower.

Accordingly, the entire concept of a single user working on a single CPU with its own memory and storage is quickly being expanded beyond the box in response to the kinds of complex problems mentioned earlier, particularly with BIM. This is the impetus behind Autodesk's A360 series of large-scale distributed computing and collaboration offerings, such as BIM 360 Team / Collaboration for Revit, Cloud Rendering, Insight 360 building energy analysis, ReCap 360, BIM 360 Docs, BIM 360 Field and Glue, Structural Analysis for Revit, FormIt 360, Building Ops, and the list goes on.

Today you can readily tap into distributed computing cycles as you need them to get a very large job done instead of trying to throw more hardware at it locally. For example, you could have a series of still renders that need to get out the door, or a long animation whose production would normally sink your local workstation or in-house Backburner render farm. Autodesk's Cloud Rendering service almost immediately provided a huge productivity boon to design firms, because it reduced the cost of getting high quality renderings from hours to just a few minutes. Autodesk isn't the only one doing this - there are hundreds of smaller dedicated render farm companies which will provide near-zero setup of dozens of high-performance CPU+GPU combinations to get the job done quickly and affordably.

Similarly, the problem of wide-area design collaboration and coordination is neatly handled by the combination of BIM 360 Team and Collaboration for Revit (also called C4R). BIM 360 Team gets you basic access to the Team Hub, where your data is stored and projects are managed. C4R is an add-on for Revit that adds the collaborative component to the A360 Team license[9]. Together, this cloud-based solution removes the local-server architecture problem that all design firms share. You no longer need to routinely send models to project team members; it all lives in the cloud all of the time, and you work off of those cloud-based models in Revit. For the user, not much is different from using standard Worksharing.

[9] https://blog.microsolresources.com/2016/01/26/a-comprehensive-guide-to-setting-up-a360-collaboration-for-revit/

Building a DIY Cloud with Amazon EC2 and S3 Services

You can build out your own cloud solution. Amazon's Web services (AWS) provides pay-as-you-go computing resources in the cloud. It does this primarily through two components, Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3)

The EC2 service provides compute capacity (through virtual servers) in the cloud. This provides the ability to build a temporary virtual rendering farm for very little money, starting at about $0.65 cents per core hour for a GPU+CPU configuration. Once signed up, you have a whole host of machines at your disposal to chew on whatever problem you need solved. This is exactly what Autodesk did, as their Cloud Rendering service runs on Amazon's EC2.

Amazon's S3 storage service platform is used solely to store and serve up data. Similar to the story with EC2, Autodesk stores your Collaboration for Revit models in Amazon's S3 data center in Virginia.

Interesting note: At AU 2014 I assisted in a Families lab taught by Paul Aubin. As part of an experiment, Autodesk worked with Amazon EC2 engineers to set up the lab to be run with Revit hosted on the EC2 platform. None of the 80 or so workstations in the class were doing any of the heavy lifting at all; they were simply connected to the cloud-based computing machines which actually ran Revit. Overall, the labs ran with very few issues. Models could be easily worked on and displayed smoothly in 3D.

Thus, perhaps the future is to move toward small, low-powered machines that easily access the computing iron held in the Cloud. This is a much more robust model than VDI, because data processing happens across thousands of CPUs housed in the datacenter. Logically this leads to the point where your desktop computing power (as well as this class) will be largely irrelevant, as your PC harnesses cloud-based CPU cycles for all computing and not just when the need arises due to insufficient local resources.

Price vs. Performance Compression

One of the side effects of steadily increasing computing power is the market-driven compression of prices. Typically, the pricing differences between any two similar components of different capacities or speeds has shrunk dramatically, making the higher end option often the more logical buy. For example, today a high quality 1TB drive is about $50, a 2TB drive is about $70, and a 4TB drive is about $120. The 4TB drive nets you 4x the storage for well under three times the price.

For system memory it is a similar case. Today's 8GB DIMMs have dropped to about $4.50/GB, which is much less than 4GB DIMMs at $5.65/GB. With modern CPUs supporting more RAM, there is an argument to be made to simply go for 32GB as (2) 16GB modules from the start. This changes a bit under the new DDR4 RAM specification, as we will see later in our section on RAM, but the idea is the same.

Processors are a different story, however. CPU pricing is based upon capability and popularity, but price curves are anything but linear and actually work in the opposite direction. A 3.2GHz CPU might be $220 and a 3.4GHz is incrementally higher at $300, but going with a 3.5GHz CPU – which is only 100MHz faster and nothing you would ever notice on the desktop - could be a $600 upcharge. This makes for plenty of "sweet spot" targets for each kind of CPU lineup, which we discuss in great detail later in this handout.

Graphics cards are typically set to price points based on the GPU and the amount of memory on the card. Both AMD (which owns AMD) and Nvidia may debut 5 or 6 new cards a year, typically based on the latest GPU architecture. Models of a particular GPU architecture are differentiated by base clock speed, onboard memory, or number of internal GPU cores present or activated. Both companies issue reference boards that 3rd party card manufacturers use to build their offerings. The result being that prices between different manufacturer's cards using the same base model number (e.g. the GTX 1080) may only be between $0 and $20 of each other, with more expensive variations available that have game bundles, special coolers, or have been internally overclocked by the manufacturer.

III. Building Design Suite Application Demands

As discussed, within any workstation there are four primary components that are stressed to some degree by the Autodesk AEC applications, and directly determine overall performance in some key area: the processor (CPU), system memory (RAM), the graphics card (GPU), and the storage subsystem. While the chipset is an important component which glues the CPU and RAM together on the motherboard, it is not by itself stressed by applications directly. If anything, the chipset in conjunction with the CPU defines a computing platform and largely determines how expandable and capable the system is in the overall.

Different applications stress different parts of your system. Given the current state of hardware, today's typical entry-level workstation may perform well in most, but not all of the applications within the Suite, due to specific deficiencies in one or more subsystems. You need to evaluate how much time you spend in each application - and what you are doing inside of each one - and apply that performance requirement to the capabilities of each component.

Autodesk AEC Application / Hardware Demand Matrix

The following table provides a look at how each of the major AEC applications are affected by the different components and subsystems in your workstation. Each value is on a scale of 1-10, where 1 = low sensitivity / low hardware requirements and 10 = very high sensitivity / very high hardware requirements.

	CPU Speed / Multithreading	System Ram - Amount / Speed	Graphics Card GPU Capabilities	Graphics Card Memory Size	Hard Drive Speed
Revit	10 / 9	9 / 7	7	5	10
3ds Max	10 / 10	9 / 7	7 / 5 /10 (Nitrous / mr / Iray™)	6 / 10 (mr / Iray™)	10
Navisworks Simulate Navisworks Manage	6 / 6	5 / 5	5	5	6
InfraWorks 360	9 / 8	8 / 6	9	7	9
AutoCAD (2D & 3D)	6 / 6	5 / 5	5	5	6
AutoCAD Architecture / AutoCAD MEP	8	7	6	5	6
ReCap 360	10 / 10	9 / 5	8	7	10

To give a sense of direction to the above chart, any low-end workstation you could purchase off the shelf[10] will likely perform adequately well in most of these applications, giving a performance rating up to about a 6. According to the chart above, such a baseline system would be adequate for AutoCAD 2D and 3D drawing tasks for most projects. However, it would likely need specific tweaking, e.g. a CPU model upgrade or more RAM, to adequately run AutoCAD Architecture or AutoCAD MEP, and is pretty much completely inappropriate for higher-order AEC applications such as Revit, ReCap 360, InfraWorks 360, or 3ds Max.

It is not that those applications will not run in such a baseline system; but rather, that system is not optimized for them. As you can see from the chart above, most of these AEC applications have at least one aspect which requires careful consideration for a particular component.

[10] Base machine specs: Core i5 /AMD A8 series processor, 8GB RAM, built-in video, 1TB mechanical hard drive

A Word about System Requirements and Certified Hardware

Autodesk maintains a list of System Requirements for its AEC applications on its website. However, these requirements are ***minimums***, not ideal hardware configurations. They are woefully out of date to boot.

For example, the System Requirement for 3ds Max 2017 lists 4GB of RAM (8GB recommended). While you can fire up 3ds Max on a lousy 4GB system, you would not want to more than once before throwing the system down the stairs. Don't design your workstation around such horrendous specifications.

Additionally, Autodesk maintains as database of recommended and certified hardware for its AEC applications at http://autode.sk/2edWsKO. Autodesk works with Nvidia and AMD to test their GPU hardware and software drivers for compatibility and performance. Sadly, similar to its laughable system requirements page, this database here is of dubious value. It doesn't list any 2017 products, nor does it test Windows 10. The list of certified hardware is very out of date as well, and does not list any video card under two years old. The one redeeming feature of the site is that for the graphics cards that do make the list, it provides download links to the certified drivers. That can be invaluable for correcting a newer but buggy video driver, which may be automatically installed by Windows Update without your knowledge.

Application Notes: Revit

The **Autodesk Revit** platform stresses most major components in a workstation because of the nature of the datasets you work on in building design. Most Suite users spend more time in Revit than most other applications, so tuning your workstation specifically for this application is a smart choice.

Because of the size and complexity of most BIM modeling efforts, it requires a fast CPU, a decent amount of system RAM, and a fast storage system. While it historically had rather mundane graphics demands and you could get by with pedestrian cards, Revit will make use of fast GPUs, particularly in later releases.

Let's see how each component is specifically affected by Revit:

Processor (CPU): Revit is, at its heart, a database management system (DBMS) application. As such, it takes advantage of certain technical efficiencies in modern high-end CPUs, such as multiple cores and larger internal L1, L2, and L3 high-speed memory caches. Modern CPUs within the same microarchitecture lineup have similar multiple cores and perhaps the same L1/L2/L3 caches, with the differences limited primarily to core clock speed. Differentiations in cache size and number of cores appear between the major lines of any given microarchitecture. This is particularly evident for high-end desktop CPUs or CPUs specifically designed for database servers, which have more cores per CPU, allow for multiple physical CPU installations, and vastly increased L1/L2/L3 cache sizes.

Revit's high computing requirements are primarily due to the fact that it has to track every element and family instance as well as the relationships between all of those elements at all times. Revit is all about relationships: its Parametric Change Engine works within the framework of model 2D and 3D geometry, parameters, constraints of various types, and hosted and hosting elements that understand their place in the building and allow the required flexibility. All of these aspects of the model must respond to changes properly and update all downstream dependencies immediately.

Revit requires a fast CPU because all of this work is computationally expensive. There are no shortcuts to be had; it has to do everything by the numbers to ensure model fidelity. This is particularly noticeable when performing a Synchronize with Central (SWC) operation, as Revit first saves the local file, pulls down any model changes from the Central Model, integrates them with any local changes, validates everything, and sends the composite data back to the server.

Revit is also increasingly multithreaded, in that more and more operations can be computed separately and simultaneously with other operations. According to Autodesk's Revit 2017 Model Performance Technical Note[11], Revit 2017 supports multithreading in certain operations:

- Vector printing
- Vector Export such as DWG and DWF
- Rendering. Autodesk Raytracer (ART) in particular overcomes the 16-core limitation in mental ray.
- Wall Join representation in plan and section views
- Loading elements into memory, reducing view open times when elements are initially displayed
- Parallel computation of silhouette edges, accelerating navigation of perspective 3D views
- Translation of high level graphical representation of model elements and annotations into display lists optimized for graphics cards, engaged when opening views or changing view properties
- File Open and Save
- Point Cloud Data Display
- Color fill calculations are processed in the background on another process
- Calculations of structural connection geometry is processed in the background on another process

All modern CPUs are 64-bit and meet or exceed the minimum recommended standard established by Autodesk. But with everything else, you want to choose a CPU with the very latest microarchitecture platform, the most cores you can afford, the fastest core clock speed, and the most L2 cache available. We will discuss these specific options in the Processor section of this handout.

System Memory (RAM): The need to compute all of these relational dependencies is only part of the problem. Memory size is another sensitive aspect of Revit performance. The current rule of thumb is that Revit consumes 20 times the model file size in memory, meaning a 100MB model will consume 2GB of system memory before you do anything to it. If you link large models together or perform an un-optimized rendering operation, your memory subsystem can be a key bottleneck in performance.

The more open views, you have the higher the memory consumption for the Revit.exe process. Additionally, changes to the model will be updated in any open view that would be affected, so close out of all hidden views when possible and before making major changes.

With operating systems getting more complex and RAM being so inexpensive, 16GB (as 2x8GB) is today's minimum recommended for the general professional level. 32GB or more would be appropriate for systems that do a lot of rendering or work in other AEC applications simultaneously.

Graphics: With Revit we have a comprehensive 2D and 3D design environment which requires decent performance graphics capabilities to use effectively. However, I have found that Revit performs adequately well on most projects under relatively mainstream graphics cards between $200 and $350, regardless of manufacturer.

This is mostly because Revit views typically contain only a subset of the total project geometry for the sake of clarity. Even in 3D views, one typically filters out and limits the amount of data which enables the system to respond quickly enough for most GPUs can handle with aplomb.

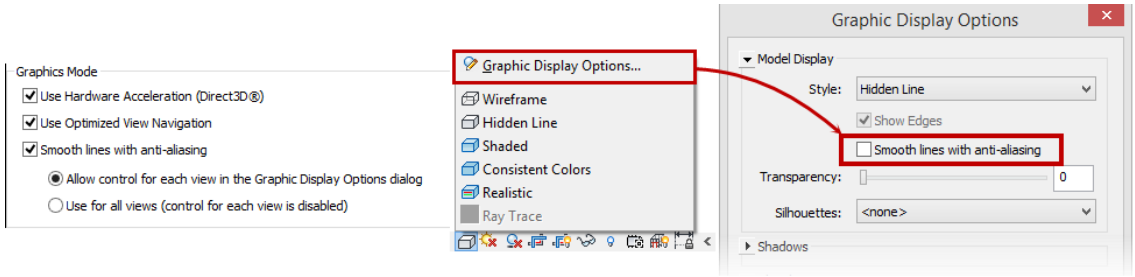
But the graphics card can get a real workout as we demand more use of shaded and realistic views complete with material appearances. Toss in sketchy lines, anti-aliasing, ambient shadows, lighting, and so on, and view performance can slow down dramatically. The better the graphics card, the more eye candy can be turned on and performance levels can remain high.

[11] http://revit.downloads.autodesk.com/download/2017RVT_RTM/Docs/InProd/Autodesk_Revit_2017_Model_Performance_Technical_Note.pdf

Your graphics performance penalties grow as the complexity of the view grows, but Autodesk is actively working to help alleviate viewport performance bottlenecks. Much of this is "low hanging fruit" from a programming standpoint, due to improvements in the Direct3D API that is in Windows (and largely driven by the video gaming industry, which outperforms the movie industry by a large degree).

In 2014, Revit viewports got a nice bump with the inclusion of a new adaptive degradation feature called Optimized View Navigation. This allows Revit to reduce the amount of information drawn during pan, zoom and orbit operations and thus improve view responsiveness and performance.

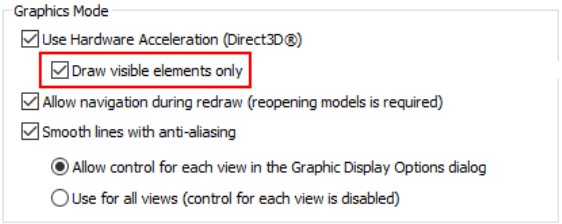
In 2015 we gained the ability to limit smoothing / anti-aliasing operations on a per-view setting using the Graphics Display Options dialog. Anti-aliasing (AO) is the technology that eliminates jagged pixels on diagonal geometry by blending the line pixels with the background. It looks great but can be computationally expensive depending on your video hardware, so view performance is optimized by turning it on only in the views that require it. These settings are found in the Options > Graphics tab and in the view's Graphic Display Options:



Revit 2015 improved performance in the Ray Trace interactive rendering visual style, and in other views, it improved drawing performance such that many elements are drawn simultaneously in larger batches using fewer drawing calls. Specifically, the 2015 release added some new enhancements to increase GPU use, namely drawing multiple geometries of the same material to the GPU as a batch, and drawing multiple identical objects to the GPU as a batch.

In Revit 2016, Autodesk has made even more strides to improve the efficiency of navigating around the viewport without waiting for it to finish drawing the model. It makes even more use of GPU processing for drawing all graphics via DirectX 11, giving Revit 2016 approximately a 30% increased drawing operation speed. Additionally, a newer, faster process is used for displaying selected objects, and the underlying technology used for displaying MEP elements in views improves performance when opening and manipulating views with many MEP elements.

Revit 2017 brings behind the scenes optimizations and improvements to over 100 functions. Navigation of 3D models is much faster with the "Draw Visible Elements Only" option, which enables an occlusion culling feature that happens at the GPU level. In other words, if you can't see it, Revit won't draw it.



All cards manufactured over the past five years will support Revit 2017's minimum system requirements[12] which specify a DirectX 11 / Shader Model 3 capable card running under Windows 7 SP1 64-bit and higher. This allows for all viewport display modes, adaptive degradation, ambient occlusion effects, and so on.

[12] https://knowledge.autodesk.com/support/revit-products/troubleshooting/caas/sfdcarticles/sfdcarticles/System-requirements-for-Autodesk-Revit-2017-products.html

The general rule is that the faster (and more expensive) the card is, the better it will be for Revit, but only up to a point. We discuss this in the section dedicated to graphics.

However, it is important to check the Graphics tab in the Options dialog to ascertain the compatibility of your particular combination of video card and driver. While Revit has often been largely video hardware agnostic in the past, this is changing. Lately I have seen many cases of video driver instability with certain driver combinations, sometimes requiring me to install much older drivers to get Revit to be stable. Especially with newer hardware, Revit will likely report "Unknown video card" as Autodesk has not tested Revit under your combination of hardware and driver. If you have to turn off hardware acceleration to get Revit to function, test older drivers to see if one works. If that does not work, get a different video card.

Storage: It is no secret that the size of the files Revit creates are huge, particularly compared to traditional AutoCAD DWG files and represents a bottleneck in opening and saving projects. 60MB Revit .RVT files are typical minimums for smaller projects, with file sizes of 100MB+ more likely. MEP models typically start around 60-80MB for complete basic projects and go up from there. Today, models topping 1GB are not uncommon.

For Workshared projects, Revit needs to first copy these files off of the network to the local drive to create your Local File, and keep that file synchronized with the Central Model. While we cannot do much on the network side without moving to more exotic infrastructures - we are all on Gigabit LANs these days, which is the most common network speed - file I/O can take their toll on your local storage subsystem.

Also, pay very close attention to your %TEMP% folder, located under the AppData\Local\Temp folder under your Windows user profile folder (C:\Users\{Username]\). When you launch Revit, even with loading small project files, your %TEMP% folder fills up with temporary Revit files. While these temporary files should erase themselves when Revit closes out, that can build up over time particularly if Revit crashes and leaves them behind. Make it a point to clean out your %TEMP% folder on a weekly basis (and reboot your machine regularly) especially if you use a smaller (<500GB) Solid State Drive.

Also, don't forget that Revit itself is a large program and takes a while just to fire up, so you want a fast storage subsystem to comfortably use the application with large models. Revit is certainly an application where Solid State Drives (SSDs) shine.

Modeling Efficiently is Key

Overall, Revit performance and model size is as much directly tied to implementing efficient Best Practices in your company as it is the hardware. A badly executed 200MB model will perform much worse than a solidly built 300MB model. With such inefficient models, Revit can consume a lot of processing power in resolving things that it otherwise would not need to deal with.

Best practices dictate efficient workflows. Create families with their 3D elements turned off in 2D views, and use fast Symbolic Lines to represent the geometry in plan and elevation instead. This minimizes the amount of information Revit will need to process in performing the hidden line mode for 2D views. In 3D views, use the Section Box tool to crop the area to minimize the number of polygons on screen. The use of Filters to turn off large swaths of unnecessary geometry can be a huge performance boon, particularly in the MEP disciplines where you can have lots of stuff on screen at one time.

Application Notes: 3ds Max

Autodesk 3ds Max has base system requirements that are about the same as they are for Revit. However, 3ds Max stresses your workstation differently and exposes weakness in certain components. Your workflows dictate how much Max has to work.

3ds Max is primarily all about having high end graphics capabilities that can handle the display and navigation of millions of polygons as well as large complicated textures and lighting. For typical AEC imagery and animations, the main areas that Max deals with are:

- Polygons - Interacting with millions of vertices, edges, faces, and elements on screen at any time;
- Materials - Handling physical properties, bitmaps, reactions to incoming light energy, surface mapping on polygonal surfaces, and procedural texture generation;
- Lighting - Calculating physical and non-physical lighting models, direct and indirect illumination, shadows, reflections, and caustics;
- Rendering - Combining polygons, materials, lighting, and environmental properties together to produce final photorealistic imagery; ray tracing under various Autodesk and 3rd-party rendering engines; and performing post-rendering effects;
- General computation - Linking RVT and FBX files adds quite a bit of computational overhead, as does working with subobject animation (e.g., kinetic sculptures, flying birds, etc.).

Each system component affects performance as follows:

CPU: 3ds Max is a highly tuned and optimized multi-threaded application. Geometry, viewport, lighting, materials, and rendering subsystems can all be computationally expensive and 3ds Max will take full advantage of multiple cores / processors. Even the UI is largely multithreaded, so having many cores at your disposal allows for fast interaction with the program even with very large and complex scenes.

As discussed, the standard scanline, mental ray, and newer Autodesk Raytracer (ART) rendering engines are almost wholly CPU dependent and are designed from the ground up to take advantage of multiple processors. Rendering times scale pretty linearly with your CPUs capabilities; as with general interactivity with the program, having one (more more) physical CPUs with multiple cores will shorten rendering times considerably. In addition, Max includes distributed bucket rendering with Backburner, which allow you to spread a single rendering task across physical machines, even further reducing rendering times.

The rule of thumb is that, even more so than Revit, 3ds Max can make whole use of the best CPU you can afford. If you spend a lot of time in 3ds Max and regularly render high resolution images, you need to look at powerful workstations that feature a high number of cores on the CPU. The Return on Investment (ROI) for high end hardware is typically shorter for Max than any other Autodesk AEC application.

CPU Limitations: An interesting note is that 3ds Max – as well as other Autodesk applications - may not be able to use all of the processor cores in your system, due to a "gotcha" of sorts with Windows Processor Groups. In short, Window will group CPU cores into Windows Processor Groups, up to 64 cores per group. Some newer monster CPUs have 18 cores, and with Hyperthreading you get 36 cores. If you put two of these physical CPUs in a single system you get 72 cores, which is more than 64 and thus get split into two Windows Processor Groups of 36 cores apiece. The problem is that 3ds Max is not "Windows Processor Group aware" and so will only see one group - 36 cores total.[13]

Also: Rendering is a huge processor-intensive aspect of 3ds Max and is probably the main reason why one would buy a 72-core machine in the first place. However, your rendering engine of choice may (or may not) suffer this same CPU core limitation. Consult your renderer documentation for details.

[13] http://area.autodesk.com/blogs/max-station/n268-how-many-cores-does-3ds-max-support

RAM: 3ds Max also requires a good deal of system memory, particularly for large complex scenes with Revit links as well as for rendering operations. The application itself will consume about 740MB with just an empty scene and minimal plug-ins loaded. If you regularly deal with large animation projects with complex models and lots of textures, you will want to specify more RAM. As we discuss in the section on CPUs and chipsets, the choice of CPU decides how much RAM your system can address. Mainstream desktop CPUs support only up to 32GB, but higher-end processors can easily support 64GB, 128GB, and on up to 1.54TB on *very* high end models.

Most AEC scenes can readily work in 3ds Max with 32GB to 64GB, after which there is perhaps a case of diminishing returns. However, for those who regularly work with large complex scenes and have multiple applications open at once, moving to a more aggressive hardware platform with multiple physical CPUs will, as a side benefit, result in more addressable RAM and provide that double benefit to the Max user.

Note that these CPU and RAM guidelines are the same for any machine used in a rendering farm as well; rendering jobs sent to non-production machines with a low amount of RAM can often fail. The best bet is to ensure all machines on an in-house rendering farm have the required amount of RAM installed and, as much as possible, the same basic CPU capabilities as your primary 3ds Max machine.

As with CPU cores, the various renderers used in 3ds Max may have different requirements for RAM and some may have specific settings for tweaking their use of system / virtual memory.

Graphics: With 3ds Max 2017 we have a continually improving viewport display system (Nitrous) which is working to take more direct advantage of the GPU capabilities in various ways. The Nitrous viewport allows for a more interactive, real-time working environment with lighting and shadows, which requires higher-end graphics hardware to use effectively.

In 2014 Nitrous got a nice bump in viewport performance with support for highly complex scenes with millions of polygons, better depth of field, and adaptive degradation controls that allow scene manipulation with higher interactivity. In 2015 viewports were made faster with a number of improvements accelerating navigation, selection and viewport texture baking. Anti-aliasing could be enabled with minimal impact on performance but real-world experience says this largely depends on the graphics card. Nitrous improvement continue in 3ds Max 2016, where users report much better viewport performance.

In 3ds Max 2017, the user interface has a new look and feel, and it properly scales up to the latest high DPI (4K) displays, something almost no other Autodesk application does. Additionally, you can dial in the exact performance you are looking for in the Viewport Setting and Preference dialog box.

In the case of Max (and, to a lesser extent, Revit), the amount of onboard video RAM (VRAM) on the graphics card can play a big part in viewport and rendering performance. Video RAM is very fast temporary storage for the GPU. It stores textures, shadow / lighting maps, and serves as a frame buffer for the entire screen. Thus, viewports that are configured for Realistic mode with shadows and lighting turned on will consume much more VRAM than simple hidden line viewports. If you enable anti-aliasing to smooth out the stair-step pattern at the edges of surfaces, the amount of VRAM required grows considerably.

A frame buffer essentially is the entire screen image stored as a large bitmap, so the resolution of your screens determines the amount of VRAM used for that. The frame buffer stores 24 bits per pixel (8 bits each for red, green and blue) plus another 8 bits per pixel for an alpha channel to store transparency information. So, at 32 total bits per pixel, a 1080p screen with a resolution of 1920 x 1080 = 2,073,600 pixels, which requires about 8MB per buffer. If you have two screens, that doubles to 16MB. Today's 4K screens offer twice the resolution, but at 3840 x 2160 this is actually 4x the pixels = 32MB per buffer.

Applications such as games will render many screen passes in quick succession, storing several frames in VRAM in order to composite them into the final image, further increasing VRAM requirements. So much so that cards with 4GB are considered low end, with higher-end cards shipping with 8 or 12 GB of VRAM.

Another consideration that may affect your choice of graphics card is the rendering engine used. Unlike mental ray or ART, the Iray rendering system can directly use Nvidia GPUs to accelerate rendering tasks to a very high degree. In addition, Iray performance scales directly with the number of GPUs in your system. Additionally, Iray requires the entire image to be stored in VRAM, which may push you towards the higher end cards with 8-12GB as well.

Oddly, your choice of Iray also obliquely plays into the choice of CPU /chipset platform. Because Iray performance scales with the GPU cores, heavy Iray users often stuff more than one graphics card into a system. The CPU platform determines the number of PCI Express slots, so if you want 3, 4, or even 5 graphics cards to leverage in Iray productively, you necessarily need to specify a CPU hardware platform that can handle multiple graphics cards at high speeds. We specifically discuss the needs of Iray users in 3ds Max in the section on graphics hardware.

Lastly, 3ds Max users on Subscription can now enjoy A360 cloud rendering support, allowing machines with minimal CPU or GPU resources to quickly render images without tying up their desktop or configuring rendering farms.

Storage: The 3ds Max program itself can be notoriously slow to load, particularly if you use a lot of plugins. Factor in the large .max files you create (particularly if you link Revit files), a fast local storage system will pay off greatly.

Finally, remember that Viz Wizards will often work simultaneously in other programs, such as Photoshop, Mudbox, Revit, Inventor, and AutoCAD, so make sure your workstation specification can cover all of these bases concurrently.

Application Notes: Navisworks Manage

Autodesk Navisworks Manage is primarily used by the construction industry to review, verify, and simulate the constructability of a project. Its three main features are the Clash that identifies and tracks collisions between building elements before they are built; the TimeLiner which applies a construction schedule to the building elements, allowing you to simulate the construction process; and integrated 2D and 3D quantification for performing easy takeoffs for estimating purposes.

The heart of Navisworks is a 3D model engine that is quite a bit different from that in Revit or 3ds Max. Navisworks does not create geometry itself; rather, you import geometry from any number of external applications (I believe Navisworks supports upwards of 30 file types). In most cases, source models may be built from scratch or start as design intent models and extended by trade contractors to include fabrication information, so they may contain much more detail than those from Revit, Max, or SketchUp.

However, when taking in a complex 3D model, Navisworks keeps the data behind the geometry but breaks the model itself down into simpler “shell” geometry and automatically removes detail that is not seen, culling back faces and so on, making very lightweight files in the process.

As such, Navisworks is all about fast viewpoint processing to allow you to navigate very large and complex building models easily. Minimizing the number of polygons allows for more fluid view navigation.

One of the biggest criticisms with this process was that, while Navisworks will easily handle navigation through a 2 million SF hospital project with dozens of linked models, the graphics look bland and not at all lifelike. Realistic imagery was never intended to be Navisworks’ forte, but this is getting a lot better with each release. We have the multi-threaded Autodesk Raytracer rendering engine, cloud rendering

using the Autodesk 360 service, and improvements in using ReCap point cloud data. Viewports now disable obscured objects not seen by the camera and have improved faceting with Revit files.

Processor: Navisworks was engineered to perform well on rather modest hardware, much more so that Revit or 3ds Max. Any modern desktop processor will handle Navisworks just fine for most construction models. Larger models will demand faster processors, just as it would in Revit and 3ds Max. But because Navisworks works on much smaller and more efficient files, performance on even very large projects does not suffer in the same way.

Surprisingly, Navisworks-centric operations, such as Time Liner, Quantification, and Clash Detective, do not require a lot of horsepower to run fast. Clash tests in particular run extremely fast even on modest hardware because the geometry is simplified and the testing algorithm is very efficient. However, the new Autodesk rendering engine will demand higher performance systems to render effectively, so if you are planning to do rendering from Navisworks, target your system specifications similar to Revit and 3ds Max.

RAM: Navisworks 2017 by itself consumes a rather modest amount of RAM - about 180MB on my system without a model loaded. Because the .NWC files it uses are rather small, additional memory required with your construction models is also pretty modest. I have found that standard 8GB systems will still work well with Navisworks on moderately sized projects, but I would encourage 16GB as a prudent minimum.

Graphics: The geometric simplification from the source CAD/BIM files allows for more complex models to be on screen and navigated in real time. In addition, Navisworks will adaptively drop out geometry as you maneuver around to maintain a minimum frame rate, so the better your video subsystem the less drop out should occur. Since there are far fewer polygons on screen, Navisworks won't test your graphics card's abilities as much as other applications. Most decent cards that would be applicable for Revit and other Autodesk's AEC applications will handle moderately complex Navisworks models without issue.

Storage: The files Navisworks creates and works with (.NWC) are a fraction of the size of the originating Revit/CAD files. NWCs only store the compressed geometry of the original application file and the source BIM data, but strips out all of the application specific data it does not need, e.g. constraints. A 60MB Revit MEP file will produce a Navisworks NWC file that might be 1/10th the size. This lowers the impact on your storage and network systems, as there isn't as much data to transfer.

Bottom Line: Overall, Navisworks has some of the more modest requirements of Autodesk's AEC applications in terms of system hardware. Because most Navisworks users are Revit users as well, outfitting a workstation suitable for Revit will cover Navisworks' needs just fine.

Application Notes: Autodesk Recap 360 and Recap 360 Pro

Autodesk ReCap 360 and 360 Pro is reality capture processing software that allows you to import, index, convert, navigate, and edit raw 3D scan data. It also includes a photogrammetry component, where you can import photographs and video, converting them to point cloud data as well. ReCap 360 saves this indexed cloud data to the highly efficient .RCS file format which can then be linked into AutoCAD, Revit, Navisworks, and 3ds. Once linked into a design application, you can snap to and trace the points in the cloud file to recreate the geometry to be used downstream.

Processor: Probably the single most expensive operation for your CPU is going to be in the indexing of raw point cloud scan files into the .RCS format. The indexing operation is heavily reliant on the CPU, RAM, and disk systems, and CPU utilization can easily peg 100% which will tank performance elsewhere. Having a very fast modern processor with many cores will definitely make the index operation faster.

Once the scans are indexed and in ReCap 360, however, CPU utilization goes down quite a bit. A heavy test project consisting of 80 .RCS files (about 18GB total) was not a problem for a workstation with only

8GB of RAM. Typical operations, such as cropping point cloud data, turning individual scans on and off, and so on were fairly straightforward without an excessive performance hit.

Memory: ReCap 360's memory consumption by itself is pretty lightweight, around 150MB. When indexing point cloud scans, RAM consumption will jump to between 500MB and 1GB. Loaded up with 18GB of test .RCS files, memory consumption only hit about 900MB, demonstrating the effectiveness of the indexing operation. Modestly equipped workstations will probably handle most ReCap projects without issue.

Graphics: The ability to navigate and explore point clouds in real time is very compelling - it's like walking through a fuzzy 3D photograph. To do this effectively means you need a decently powered graphics card, as ReCap 360 has to display billions of points and navigate through them in real time. A marginal system without a fast GPU will definitely suffer in display responsiveness no matter how small the project.

Storage: A typical scan project may have many large (100-300MB) individual point cloud .RCS scan files, so a ReCap 360 project of 50 or so scans will consume many GB of disk space. With such large datasets, Solid State Drives (SSDs) will definitely help ReCap 360's core operations as it can work on that volume of data very quickly. Disk requirements may also impact your server's storage capabilities as well.

However, ReCap 360 is, like all "360" applications, cloud-based, so you can upload, store, and edit scan and photographic data in the Cloud instead of using local resources. This may alleviate some of the pain you may experience while using marginal local hardware resources.

Application Notes: AutoCAD / AutoCAD Architecture / AutoCAD MEP

Autodesk AutoCAD 2017, having been around for so long, has hardware requirements that are pretty well understood and can be handled by modest entry level workstations. Although new releases may add new features here and there, for 2D drafting and design, any modern PC or workstation should suffice. For vertical applications - AutoCAD Architecture (ACA) and AutoCAD MEP (AMEP) - hardware requirements go up (albeit slightly) because of their code complexity as well as the increased use of 3D geometry.

Processor: All modern CPUs will largely handle AutoCAD, ACA, and AMEP tasks without issue. As your projects get larger and you work with more AEC objects, CPU usage will climb as AutoCAD Architecture and MEP needs to calculate wall joins, track systems, schedule counts through external references, and other more CPU intensive operations. But it's all still child's play from a CPU workload perspective.

System Memory: Most systems with equipped with 8 to 16GB will handle base AutoCAD and its vertical just fine. In informal testing, AutoCAD 2017 consumes about 200MB by itself without any drawing files loaded (although this depends on what external add-ins and applications, such as Visual LISP/ActiveX code, is loaded on Startup). ACA 2017 weighs in at 400 MB, and AMEP at 420MB with standard templates. AutoCAD's verticals will consume a lot more memory because of the additional AEC specific information held in each object, the code added to the DWG file, and in handling display configurations. Regardless of flavor you are using, RAM increases as you have more drawing files open, have many layout tabs in each file, and have layouts with many viewports.

Graphics: The needs of 2D CAD have been well handled by moderately priced graphics cards for some time. However, for 3D CAD, ACA and AMEP work, a higher-end graphics card will pay off with faster 3D operations such as hide, orbit, and display representation operations. If you only do 2D CAD in AutoCAD but also do 3D work in other Suite programs like 3ds Max, ensure your graphics capabilities can adequately match the higher demand of those higher-order applications.

Storage: All AutoCAD based applications work with comparatively small .DWG files, so storage requirements are easily met on baseline systems. As with all AEC applications, AutoCAD and particularly the verticals can take a long time to load, and thus will benefit from fast disk subsystems in that regard.

IV. Processors, Chipsets, and Platforms

Introduction

Selecting a central processing unit (CPU) is part and parcel part of defining what ***computing platform*** will be used, and is all about comparing specifications and relative costs. There are currently four primary computing platforms that you would consider: The ***Mainstream Desktop***; the ***High-End Desktop***; the ***Classic Workstation***; and the **Mobile Workstation**. Each are differentiated in the primary type of CPU they use and their relative capabilities as compared to one another. These platform choices make up a large part of this class and are discussed in great detail in this handout.

To understand any single platform and its benefits, one must understand the value of its constituent components. Platforms are primarily defined by the processor and the ***chipset***. A chipset is an integrated circuit of electronic components that manages the data flow between the processor, memory, and system peripherals. It is an integrated part of the motherboard and designed to work with a specific family of microprocessors. It is literally the glue that binds the CPU to the rest of the machine.

Once you decide on particular platform, you are then left to choose between one or more available CPU models of the same basic product class. CPUs in the same product class may only differ by minor properties, such as the system core speed or amount of onboard cache, but this can have a huge impact on price. Two CPUs of the same platform class may differ in core speed by only 100MHz - which is inconsequential on a 3+ GHz processor - but may differ in cost by hundreds of dollars.

The microarchitecture of the chip and the process by which it was made advances year after year, so your attention will naturally focus on the latest and greatest models when specifying a workstation. This section will discuss four primary kinds of Intel CPUs (one for each platform): The latest 5th-generation "Skylake" line of mainstream desktop CPUs, the Broadwell E "High End Desktop" (HEDT) lineup, the Workstation-class 4th generation Broadwell EP Xeon E3 / E5 v4 lineup, and the latest 6th & 7th generation Core i7 mobile processors. Along the way we'll discuss how Intel develops CPUs over time, and what each kind of CPU brings to the table. All of the technical information for Intel CPUs can easily be compared at http://ark.intel.com.

We will also look at other deciding factors - chipsets, memory architectures, and expansion capabilities that will segregate workstation classes for the three user profile types discussed at the end of Section I.

Intel or AMD?

Note that this class focuses only on Intel processors. AMD has made x86 compatible CPUs or a long time, but in recent years have never competed well against the kinds of high-end performing offerings from Intel, and subsequently have been pushed out of the AEC user market. AMD CPUs generally do well at the low end of the scale, but even there have a tough time competing with Intel on speed, value, and especially power consumption.

However, this may be changing with AMD's new "Zen" architecture, expected for release in early 2017. As a clean-sheet design that differs greatly from the long-standing but ineffective Bulldozer architecture, Zen-based processors will use a 14 nm FinFET process (the same as Intel's latest and greatest Skylake architecture), are reportedly more energy efficient, and have a significantly higher instructions per cycle. Simultaneous multithreading (AMD's answer to Intel's Hyperthreading) has been introduced, allowing each core to run 2 threads. The cache system has also been redesigned, making the L1 cache write-back. Additionally, Zen based processors will bring Intel-only features such as DDR4 and PCIe 3.0 support.

All signs look good that AMD can finally bring its CPUs in line, at least performance-wise, with Intel's Core i7 desktop and HEDT processors.

That's a huge deal, since AMD has historically lagged behind Intel in the medium- to high-end desktop space. Early benchmarks show a Zen CPU trading top scores with a Broadwell E CPU. However, actual shipping unit performance benchmarks and pricing structures remain to be evaluated.

If Zen is a hit, it should force two things to happen. First, Intel's de facto monopoly on the desktop and high-end desktop market may be diminished and prices, especially at the high end, should fall to remain competitive. Second, the impetus to keep ahead of AMD should spur Intel to create better, faster CPUs with each iteration and to push forward with the next step, the 10nm fabrication node.

Until then, however, Intel's current offerings will have to do and are the focus of this section.

Intel's Microarchitectures and Processes

Before we talk about the specifics in today's CPU models and platforms, we should discuss how Intel develops their chips. This will let you understand what's under the hood when making processor and platform choices.

First some definitions: The term **"microarchitecture"** refers to the computer organization of a particular microprocessor model. It is defined as "the way a given instruction set architecture is implemented on a processor[14]." Microarchitectures describe the overall data pipeline and the interconnections between the components of the processor, such as registers, gates, caches, arithmetic logic units, and larger elements such as entire graphics cores. The microarchitecture decides how fast or slow data will flow through its pipeline and how efficient that pipeline runs. Microprocessor engineers are always looking to ensure no part of the CPU is left unused for any length of time; an empty pipeline means that data somewhere is waiting to be processed and precious cycles are being wasted as nothing gets done.

Every release of a new microarchitecture is given a code name. From the 286 onward we've had the i386, Pentium P5, P6 (Pentium Pro), NetBurst (Pentium 4), Intel Core, Nehalem (Core i3, i5, i7), Sandy Bridge, Haswell, and finally Skylake in 2015. The next major microarchitecture will reportedly be called Icelake.

Within each microarchitecture we also get incremental improvements which get their own code names, such as Ivy Bridge, Broadwell, and the upcoming Kaby Lake, so keeping things straight is also a hurdle.

Note that microarchitecture is separate from the manufacturing process itself. As previously discussed, the terms **"technology node," "manufacturing process,"** or simply **"process"** primarily refers to the size of the lithography of the transistors on a CPU, and is discussed in terms of nanometers (nm). The same microarchitecture can be implemented on different processes. When a microarchitecture is developed on a smaller node it is often referred to as a **"die shrink."**

From a history perspective, over the past 10 years we've gone from a 65nm process in 2006 with the P6 and NetBurst microarchitectures (which included the Pentium 4, Pentium M and Celeron lines); down to a 45m process in 2008 with Core and Nehalem; then down to a 32nm process with Sandy Bridge in 2010; to a 22nm process with Haswell in 2012, then down to a 14nm node process with 2015's Skylake, where we are currently holding steady. While we are then expected to get to 10nm in 2017, it's a dicey speculation, after which the process crystal ball gets very cloudy indeed.[15]

CPU Caches

Back in the early days of computing, main system memory was extremely slow and incredibly expensive compared to today. In 1993 a 4 ***Mega***byte RAM module ran about $110, or $27.50/MB. Today you can get a 4 ***Giga***byte RAM module (1000x times larger) for only $21, or about half a penny per MB.

[14] http://en.wikipedia.org/wiki/Microarchitecture

[15] https://en.wikipedia.org/wiki/List_of_Intel_CPU_microarchitectures

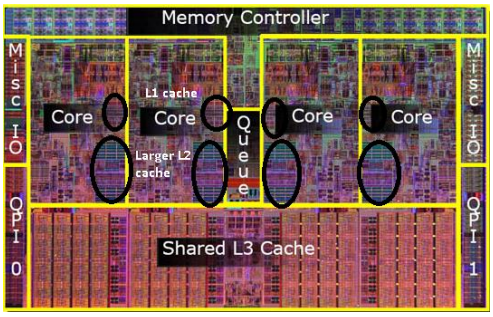
While early CPUs weren't particularly fast, the gap between CPU clock speeds and memory clock speeds began to widen very quickly from the early 1980's. Modern CPUs now run anywhere from 2 to 4 GHz and beyond, and far outstrip the ability for system memory - running around 1.6 GHz - to keep pace.

This is a problem because every time the CPU needs to recall a piece of data from system RAM it needs to stop what it is doing and wait for the RAM to feed it the information. This is called ***latency*** and was identified very early on as a serious bottleneck to throughput and performance, so CPU designs started to include internal memory pools on the die itself that can keep up with internal data requests.

These system pools are called ***Caches*** and are designed to store information that the CPU is most likely to need next. Most CPUs have different independent caches for instruction and data, where the data cache is usually organized as a hierarchy of more cache levels - Level 1, Level 2, Level 3, and in some cases Level 4, or simply L1, L2, L3 and L4.

L1 cache is the fastest, L2 is slightly slower than L1, L3 is slightly slower than L2, etc. The L1 cache is the fastest because it is built using larger transistors and wider metal tracks, and thus consumes more space and power.

This space / power penalty thus makes it the smallest capacity cache level (typically only 32KB). The L2 cache is more tightly packed than the L1 cache, and uses smaller transistors (and so there can be more of it, usually 256KB).



The L3 cache is even more tightly packed and much larger, e.g. 8MB. The sizing of each cache level affects the efficiency of the cache and overall CPU performance. Both L1 and L2 are integrated into each processor core in a multi-core CPU, another reason why they are the fastest caches. The L3 cache, on the other hand, is still on die but shared by all cores in a CPU.

Which information is loaded into which cache depends on sophisticated algorithms and certain assumptions about programming code. The goal of any cache system is to ensure that the CPU has the next bit of data it will need already loaded into cache when it goes looking for it. This is called a "cache hit," and ensuring cache hits happen regularly is a primary goal of the CPU designer.

A cache miss, on the other hand, means the CPU has to go to the next (slower) cache level to find the data. While it will be slower, it will also be much larger, hopefully maximizing the chance of a cache hit here instead of somewhere later and slower down the line. If data can't be found in the L2 cache, the CPU continues down the chain to the L3 cache - typically still on-die - then to L4 (if it exists), then finally to main system memory RAM. Of course, if the data is not in system RAM is then goes to the hard disk.

Typically the amount of L1 and L2 cache on any particular CPU microarchitecture is constant. Individual CPU models may differ – sometimes greatly - in the amount of L3 cache. The amount of L3 cache can be a very important performance factor in database applications such as Revit.

The Tick-Tock Development Model

As mentioned above, both the microarchitecture and process technologies move on different tracks. To balance the work between microarchitecture and process advancements, Intel adopted a "Tick-Tock" development strategy in 2007 for all of its future processor development cycles.

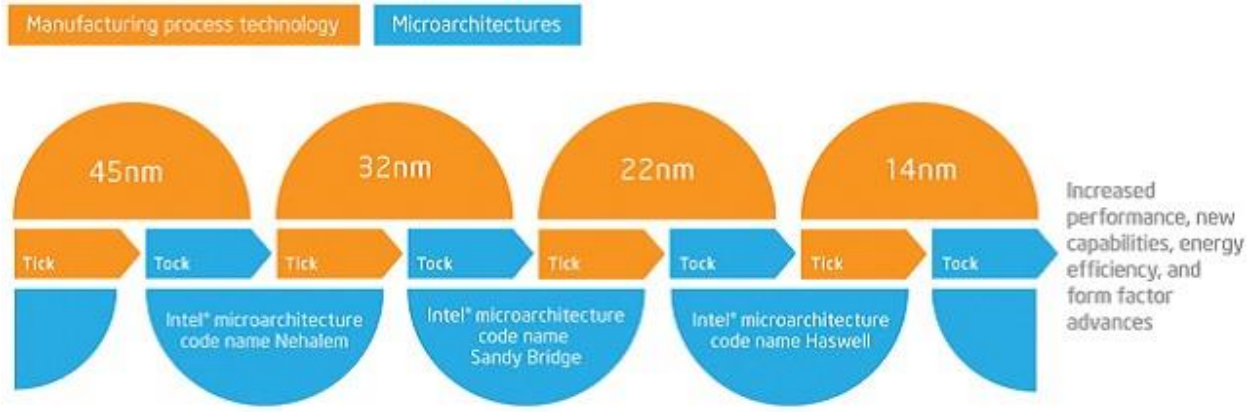
This strategy has every introduction of a new microarchitecture be followed up with a die shrink of the process technology with that same microarchitecture. In short, a ***Tick*** shrinks the process technology used in the current microarchitecture. As discussed in the Maximizing Performance per Watt discussion in Section II, shrinking a process is very hard and a big deal, because if it were easy we'd already be at the

smallest process possible. Each time Intel wants to shrink the size of the process, they pretty much has to invent tons of new lithography technology along the way to get a laser beam small enough to physically etch the silicon. Remember we are down to 14nm today with roadmaps down to 10nm and then perhaps to 7nm. The technical hurdles to go smaller and still maintain CPU cohesion and stability are enormous.

In addition to the die shrink, Ticks include small but important tweaks to the CPU cores as well, but usually nothing Earth-shattering. With a Tick you essentially get the same CPU design as last year; but with a smaller process comes lower power consumption (which then lowers heat dissipation requirements and thus lowers system noise). Along the way you get bug fixes, new instructions, internal optimizations, and slightly higher performance at the same or lower prices.

Of course, each die shrink Tick also gets a new code name as well. For example, the Westmere "tick" was a 32nm die shrink of the Nehalem microarchitecture which added several new features. Ivy Bridge was a 22nm die shrink of 32nm Sandy Bridge, and Broadwell was a 14nm die shrink of Haswell.

Conversely, a ***Tock*** is the introduction of an entirely new microarchitecture CPU design based on that now existing smaller process. This is introduced after Intel formally vets the smaller process with the previous tick and has everything working. Every year we could expect one tick or one tock, with some variations.



Intel's Tick-Tock Model. Source: Intel

Legacy Microarchitectures and CPUs: Nehalem to Haswell E

Let's look at a brief history of CPU microarchitectures over the past few years under the Tick-Tock regime, so you can understand where your current system fits into the overall landscape. Then we will dive into the current lineups in greater detail in the next sections.

1st Generation Tock: 45nm Nehalem in 2008

Starting in 2008, we had the introduction of the Nehalem microarchitecture (as a Tock), based on the 45nm process introduced the series prior. The new Core i7 lineup of CPUs were the first quad-core processors that provided a noticeably large jump in performance, mostly due to the inclusion of several key new advances in CPU design, which rolled many functions from the chipset into the CPU itself.

First, there was now a memory controller integrated on the CPU core running at full CPU speed. Nehalem CPUs also integrated a 16-lane PCIe 2.0 controller. Taken together, these integrations completely replaced the old Front Side Bus and external Northbridge memory controller hub that was part of the chipset used to communicate with system memory, the video card, and the I/O controller hub (also called the Southbridge). This bringing of external functionality onboard to increase performance closer to CPU speeds is something Intel would increase in the future.

Next, Nehalem introduced Turbo Boost, a technology that allows the chip to overclock itself on demand, typically 10-15% over the base clock. We'll look at Turbo Boost in detail later on.

Nehalem / Core i7 also reintroduced ***Hyper-Threading***, a technology first debuted in the Pentium 4 that duplicates certain sections of the processor allowing it to execute independent threads simultaneously. This effectively makes the operating system see double the number of cores available. The operating system will then schedule two threads or processes simultaneously, or allow the processor to work on other scheduled tasks if the processor core stalls due to a cache miss or its execution resources free up.

Imagine you are me whenever I go to the grocery store. Invariably, the person in front of you has to write a check, or needs someone to price check their Doritos. You are experiencing the same kind of blockages CPUs do. Hyper-Threading is effectively having another cashier opens up their lane and lets you go through. It simply makes the processor more efficient by keeping the lanes of data always moving.

Mainstream Nehalem CPUs in this era were the quad-core Bloomfield i7-9xx series and the Lynnfield i7-8xx series, which were (and still are) quite capable processors. Bloomfield CPUs were introduced first and carried a triple-channel memory controller, which meant they had to have memory installed in threes, not twos, and motherboards required six DIMM slots instead of four, making them more expensive. The lower-powered Lynnfield i7-8xx series was introduced later which had a dual-channel memory controller and motherboards were back to four DIMM slots and much less expensive to boot.

1st Generation Tick: 32nm Westmere in 2010 and Core i7/i5/i3 CPUs

Although not trumpeted as a huge deal at the time, in 2010 we had a Tick (die shrink) of Nehalem to 32nm with the Westmere architecture. Not many people remember this because it was limited to peripheral CPUs and not very many mainstream desktop models. Westmere introduced dual-core Arrandale (mobile) and Clarkdale (low-end desktop) CPUs, the six-core, triple-channel Gulftown desktop and Westmere-EP server variants, and ten-core, quad-channel Westmere-EX, typically found on high-end Xeon CPUs meant for database servers.

In addition to the Core i7 introduced in Nehalem, Westmere introduced the Core i3 and Core i5 variants, each of which targets a specific market segment. We still see these product designations today. Core i3 CPUs are typically low powered, dual-core iterations most often seen in ultraportables and very inexpensive PCs. Core i5 CPUs are generally quad-core but do not include Hyper-Threading. Neither the i3 or i5 should be of consideration to AEC users. Core i7 CPUs are both quad-core and have Hyper-Threading, and thus forms the foundation for the mainstream desktop platform. Higher-order platform CPUs are largely extensions of this basic design, in terms of more cores, more cache, etc.

However, even this once-easy Core i7/i5/i3 segregation is changing. The newest Kaby Lake 7th generation i7-7Y75 and i7-7500U mobile processors, aside from further making the case that engineers should not name processor models, now come with only two cores (4 with Hyperthreading), making it harder to distinguish CPU power based on code name alone.

2nd Generation Tock: 32nm Sandy Bridge in 2011

In 2011 things got very interesting with new microarchitecture called Sandy Bridge, based on the same 32nm process as Westmere, but with many dramatic internal improvements. Sandy Bridge represented an impressive increase in overall performance. Improvements to the L1 and L2 caches, faster memory controllers, AVX extensions, and a new integrated graphics processor (IGP) included in the CPU package made up the major features.

Sandy Bridge was important because it clearly broke away from past CPUs in terms of performance. The on-chip GPU - one of the first to be considered even marginally good - came in two flavors: Intel HD Graphics 2000 and 3000, with the latter being more powerful. This was important for the mainstream user as it finally allowed mid-size desktop PCs (not workstations you or I would buy) to forego a discrete

graphics card. Of course, AEC designers and visualization artists require decent discrete graphics far above what an IGP can provide.

Specific processor models included the Core i3-21xx dual-core; Core i5-23xx, i5-24xx, and i5-25xx quad-core; and the Core i7-26xx and i7-27xx quad-core with Hyper-Threading lines. In particular, the Core i7-2600K was an immensely popular CPU of this era, and there are still plenty of Revit and BIM workstations out there based on this chip.

Sandy Bridge E and the Rise of the HEDT Platform in 2011

In late 2011 Intel released a new "Extreme Edition" of Sandy Bridge called **Sandy Bridge E**. Neither a Tick nor a Tock, it was intended to stretch the Sandy Bridge architecture at 32nm to higher performance levels with more cores (up to 8) and more L3 cache. This introduced the new ***High-End Desktop (HEDT)*** platform and CPU lineup, which included the 4-core Core i7-3820 with 10MB of L3 cache, the 6-core Core i7-3930K with 12MB cache ($550) and the 6-core i7-3960X with 15MB cache ($1,000). This introduction of an "extreme" or high-end desktop version will carry forward with each new microarchitecture.

However, this new HEDT platform typically follows an out-of-phase cadence with that of mainstream desktop CPU development, and generally lags one generation behind. For example, HEDT's Sandy Bridge E was introduced just when the mainstream desktop got Ivy Bridge, the 22nm die-shrink of Sandy Bridge.

This faulty cadence occurs primarily because (rightly or wrongly) the HEDT market is viewed by Intel to be more of an extension of the bottom of the enterprise / server market, rather than the extending up of the mainstream desktop market. For example, the 6-core desktop Sandy Bridge E is really a die-harvested Sandy Bridge EP Xeon. While the EP-based Xeon has all 8 cores enabled, the 6-core Sandy Bridge E simply had two cores fused off.

The enterprise / server market requires stability above all else, with model updates that appear less frequently with sufficient longevity of each CPU. Thus, a server or workstation- class CPU model may hang around for a lot longer than in a desktop lineup, so buyers need to be aware that a "new" Workstation-class system may in fact ship with a very old processor.

Regardless of Intel's reasoning in market segmentation, these 6-core i7-39xx Sandy Bridge E's and Xeon E5s made excellent workstation foundations in their day. Sandy Bridge E CPUs did not include the IGP – considered useless for workstation use anyway - but did have a quad-channel memory controller that supported up to 64GB of DDR3 system RAM and provided massive memory bandwidth. A quad-channel controller meant memory has to be installed in fours to run most effectively, which requires motherboards with 8 slots. But many AEC customers have workloads which rely on large amounts of RAM.

Another plus for the emerging GPU-compute market was the inclusion of 40 PCI Express (PCIe) 3.0 lanes on the CPU, whereas normal Sandy Bridge CPUs only included 16 PCIe 2.0 lanes. This effectively allows more than one graphics card to be installed in a system and have all of them run at an adequate speed.

The PCIe 3.0 specification basically doubles the bandwidth of PCIe 2.1, where a single PCIe 3.0 8-lane x8 slot runs as fast as a PCIe 2.1 16-lane x16 slot. However, a single modern GPU is pretty tame, bandwidth wise, and you would not see much of a performance delta at all between PCIe 3.0 x8 and x16.

PCIe 3.0's additional headroom is suited very well to GPU compute as it allows more GPUs to be installed in the system without degrading all of them to the constricting x4 bandwidth. For people who needed additional graphics cards for dedicated GPU compute tasks, the lack of PCIe 3.0 became a deal breaker. See the section on PCI Express for a fuller explanation.

However, Sandy Bridge E's PCIe 3.0 was implemented before the PCIe 3.0 standard was officially ratified, meaning that it was never fully validated for that CPU. This caused technical issues where in some cases PCIe 3.0 graphics cards would default back to PCIe 2.0 speeds, such as Nvidia's Kepler series. You could sometimes force PCIe 3.0 mode on SB E in many cases, but in others you would experience instabilities.

Sandy-Bridge E was also important in that it often traded top benchmarks with the later Ivy Bridge mainstream desktop CPU, due to the addition of two cores and higher memory bandwidth, and represented a solid investment for heavy AEC software users.

3rd Generation Tick: Ivy Bridge in 2012

Hot on the trail of Sandy Bridge E, we got a Tick die-shrink of Sandy Bridge to 22nm with Ivy Bridge in April of 2012. Backwardly pin-compatible with Sandy Bridge's LGA 1155 socket, most motherboards required a simple BIOS update to upgrade the CPU.

Ivy Bridge brought some new technologies, such as the 3-dimensional "Tri-Gate" transistor (required because at this tiny scale we need to build better electron gates). We also got a 16-lane, fully validated PCIe 3.0 controller, and relatively small improvements in speed (roughly ~5-10% over Sandy Bridge), but with a remarkably lowered power draw. At this time you started to see the emergence of the "mini-ITX" form factor emerge with very small, very silent, but fully-powered and quite capable desktop designs.

The CPU's onboard Intel HD Graphics 4000 GPU was upgraded with full DirectX 11, OpenGL 3.1, and OpenCL 1.1 support. While better than the 3000, it was still not fast enough for intense gaming when compared to a discrete card, which is one reason why the graphics card market still remained so vibrant.

Overall, the HD Graphics 4000 compared to the AMD Radeon HD 5850 and Nvidia GeForce GTX 560, both respectable cards for BIM given Revit's then-fairly mundane video requirements. For real BIM and 3ds Max users, however, it was best to avoid any mention of an IGP and get one or more dedicated cards.

The Ivy Bridge lineup included the dual-core Core i3-3xxx CPUs; the quad-core Core i5-33xx, i5-34xx, and i5-35xx CPUs; and quad-core Core i7-3770K with Hyper-Threading.

HEDT: Ivy Bridge E in 2013

2013's Ivy Bridge E was the follow-up to Sandy Bridge E, using the same core as 22nm Ivy Bridge but aimed squarely at the HEDT enthusiast (and BIM and Viz user). As with SB-E it has 4 and 6 core variants, higher clock speeds, larger L3 caches, no IGP, 40 PCIe 3.0 lanes, quad-channel memory, and higher prices. It was typically billed as a desktop version of the Xeon E5.

Unlike SB E, there was no "baseball bat to the knee" handicapping here – the 6-core CPUs were truly 6 cores, not 8 cores cut down to 6. IVB E was great for workstations in that it has fully validated 40 PCIe 3.0 lanes, which meant you could easily install three or four powerful graphics cards and get at least x8 speeds.

The Ivy Bridge E lineup included three versions: At the low end we had the $320 4-core i7-4820K @ 3.7GHz, which was largely ignored. The $555 i7-4930K represented the sweet spot, with 6 cores @ 3.4GHz and 12MB of L3 cache. The $990 i7-4960X, which got you the same 6 cores as its little brother and a paltry 200MHz bump in speed to 3.8GHz, was just stupidly expensive.

One big consideration for IVB E was the cooling system required. Because of the relatively small die area - the result of 2 fewer physical cores than on SB E - you have a TDP (thermal design power) of 130W, which is the same as the first generation of Extreme Edition Nehalem i7-9xx CPUs of '08-'09. None of the IVB E CPUs shipped with air cooling, and most chose closed-loop water cooling for the silence and capability. Intel even offered a closed-loop water cooling system for the Ivy Bridge E.

4th Generation Tock - Haswell in 2013

June 2013 introduced the new **Haswell** microarchitecture for mainstream desktops. Composed of 1.6 billion transistors (compared to 1.4 billion on Ivy Bridge), and optimized for the 22nm process, the CPU was only slightly larger than Ivy Bridge, even though the graphics core grew by 25%. Internally we got improved branch prediction, improved memory controllers that allow better memory overclocking, improved floating-point and integer math performance, and overall internal pipeline efficiency as the CPU can now process up to 8 instructions per clock instead of 6 with Ivy Bridge. Workloads with larger datasets would see benefits from the larger internal buffers as well.

Haswell was interesting because, while it was clearly faster than Ivy Bridge, it got a refresh halfway through 2014 with the "Devil's Canyon" i7-4790K model, which offered an improved thermal interface material that allowed for a modest increase in clock speed to 4.0GHz, and was supported by a new 9-series chipset (Z97). Coupled with the issues with Broadwell's release date, the i7-4790K quickly became the go-to desktop CPU in 2014 through almost all of 2015.

5th Generation Tick – Broadwell in ~~2014~~ 2015

Broadwell is Intel's 14nm die shrink (tick) of the Haswell microarchitecture, and is where we really start to see Intel's relentless tick-tock strategy break down. Intel simply could not execute the 14nm process quickly enough to keep Broadwell on schedule for a widespread 2014 release. It ended up severely delayed until early 2015, which was closing in on the new Skylake microarchitecture release date. Additionally, the introduction of Broadwell did not produce the full range of desktop CPUs that could supersede Haswell, instead being relegated to mobile platforms. As such, most publications recommended users to hold off on Broadwell and wait for Skylake, which is what just about everyone did.

HEDT: Haswell E in 2014

The HEDT crowd got a nice bonus in Q2 2014 with the Haswell E high-end desktop CPU. Based on the Haswell microarchitecture under the 22nm process, Haswell E offered a nice upgrade from Ivy Bridge E and typical desktop CPUs. At this point, HEDT officially abandons all 4-core models in favor of 6 and 8 core offerings. At the "low" end we have the 6-core i7-5820K@3.3GHz with 15MB cache at $390; the 6-core I7-5930K@3.5GHz and 15MB cache at $600; and the 8-core i7-5960X@3.0GHz and 20MB cache at $1,000.

The End of Tick-Tock

Nothing lasts forever, and with the huge hurdles facing Intel in developing new processes smaller than 14nm process, they have decided to abandon the tick-tock model in 2016 moving forward. Instead, they now favor a three-step "process–architecture–optimization" model, under which three generations of processors will be produced under the same manufacturing process instead of two.

Adding a new optimization phase to the standard tick-tock model means that Intel can still produce a somewhat new product (satisfying the demand for newer / faster / whatever that we have all become accustomed to), or, in Intel's words, it allows the company to "further [optimize] our products and process technologies while meeting the yearly market cadence for product introductions."[16] In other words, Intel can continue to pull in a revenue stream from essentially an already proven product with even very minor tweaks here and there. In 2016 Intel introduced the Kaby Lake optimization to its Skylake architecture.

[16] https://en.wikipedia.org/wiki/Tick-Tock_model

Turbo Boost Technology Explained

Along with microarchitectures, processes, cache levels, Hyperthreading, and tick-tock development cycles, we should also discuss how a CPU's core clock speed functions. Clock speed used to be simple, because a CPU had only one speed and it was easy to judge performance between CPUs based solely on that speed. When comparing the clock speeds of today's CPUs, you will notice that it is no longer given as a single number, but represented as a core clock speed and a "Max Turbo" frequency.

This is because of three things. First, CPUs now have multiple cores. Each core consumes power and gives off heat, affecting how efficient the CPU as a whole can perform. Second, not all cores are fully active at all times, depending on workloads. Third, CPU cores can be internally "overclocked" or run at a rate slightly faster than their guaranteed base frequency. Intel exploits all of these properties with ***Turbo Boost***.

Intel's Turbo Boost Technology 1.0 was introduced in Nehalem processors, and improved single-threaded application performance by allowing the processor to run above its guaranteed base operating frequency by dynamically controlling the CPU's clock rate. It is activated when the operating system requests higher performance states of the processor.

The clock rate of any processor is limited by its power consumption and temperature, which is driven by the number of cores currently in use and the maximum frequency of the active cores. When the OS demands more performance and the processor is running below its power/thermal limits, a single core's clock rate can increase in regular increments of 100MHz to meet demand up to the upper Max Turbo frequency. When any of the electrical limits are reached, the clock frequency drops in 100MHz increments until it is again working within its design limits. Turbo Boost technology has multiple algorithms operating in parallel to manage current and temperature levels to maximize performance and efficiency.

Turbo specifications for a processor are noted as a/b/c/d/… notation, where each number is the maximum turbo bin for n, n-1, n-2, n-3 … n-n-1 active cores respectively, where n is the total number of cores in the processor. For a 4-core CPU, a notation of 8/8/9/10 means that with 4 or 3 cores active, the turbo bin is 8, with 2 cores active the turbo bin is 9, and with only one core active the turbo bin is 10.

These bins are multiplied by the standard increment (100MHz in Sandy Bridge and later) and added to the base clock frequency in Mhz. For example, the i7-6700 Skylake CPU has a base frequency of 3.4GHz (3400 MHz) and a max Turbo frequency of 4.0GHz. Its Turbo bins are 3/4/5/6, which breaks down as follows:

No. of cores active	No. of Turbo bin steps	Turbo Boost Calculation	Max frequency
4	3	3400 + (3 × 100) = 3400 + 300 = 3700	3.7 GHz
3	4	3400 + (4 × 100) = 3400 + 400 = 3800	3.8 GHz
2	5	3400 + (5 × 100) = 3400 + 500 = 3900	3.9 GHz
1	6	3400 + (6 × 100) = 3400 + 600 = 4000	4.0 GHz

Note: Turbo bins for specific processor models are provided at http://www.intel.com/content/www/us/en/support/processors/000005523.html?wapkw=%28cs-032279%29.

In general, normal desktop processors have low bin numbers for the first two digits because the core base clock is already set relatively high, leaving lower headroom at the top when three or four cores are active. For example, the Haswell i7-4790K has a bin of 2/3/4/4 (4.0 to 4.4Ghz), but today's Skylake i7-6700K has a bin of 0/0/0/2, meaning it has a performance range of 4.0Ghz to only 4.2Ghz.

What is important to note is that the "Max Turbo" rate you see advertised will only be reached for instances when one or at most two processor cores are active and can be boosted. For tasks which are

heavily multithreaded and require as much CPU speed as possible (e.g., rendering), all cores will be active and thus the available turbo headroom is very small, if at all.

This is why the AEC application user should concentrate more on the guaranteed core clock (worst case) speed rather than the Turbo speed when selecting a processor. While the Turbo bin number can vary slightly between models of a given series, 100MHz differences between the active core speeds within a single CPU lineup are insignificant to the end user and not anything to get excited about.

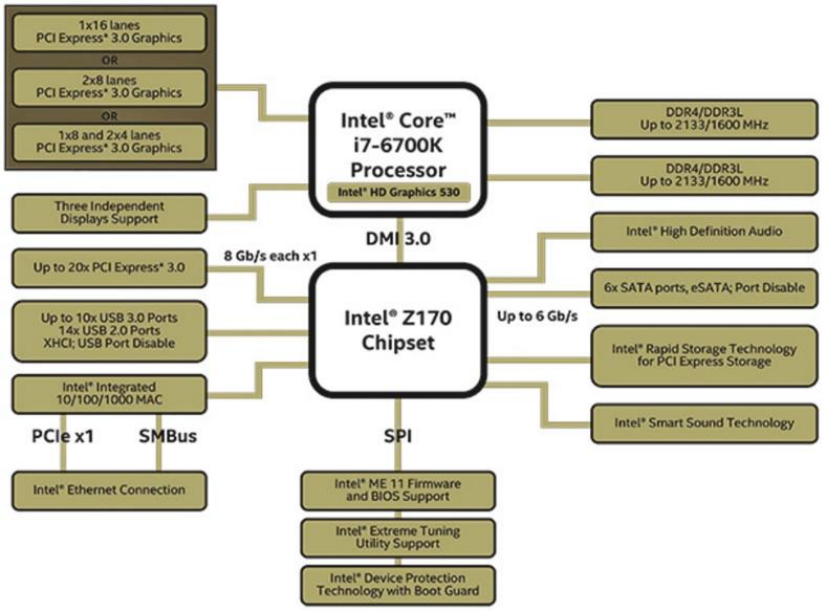
On the topic of overclocking. Intel has officially supported overclocking by users only on the enthusiast-grade K-series parts with their unlocked multipliers. It is worth noting that with Sandy Bridge and Ivy Bridge processors with "locked" (non-K) models, overclockers soon discovered that they were able to squeeze out a little more performance by increasing the Turbo bins by 4 above stock, effectively adding (at most) a 400MHz overclock.

However, this little hack was just a BIOS setting generally found in higher end enthusiast motherboards and not in something you would see from Dell or HP. From Haswell moving forward, Intel has removed all Turbo bin tweaking shenanigans, leaving the K-series as the only true overclockable processor.

Sockets and Chipsets

As mentioned in the beginning of this section, no discussion of any processor is complete without including both the microprocessor socket and the Platform Controller Hub - or **PCH**, commonly referred to as the ***chipset*** - that marries the CPU to the motherboard. For any particular CPU lineup, Intel creates a socket standard and a matching PCH to go with it, in various flavors that cater to a certain market.

Today's CPU provides four primary interfaces out of the chip: (1) Dedicated connections directly to system RAM; (2) Some number of direct PCI-Express (PCIe) connectivity for multiple graphics cards in several configurations, (3) video I/O for CPUs with an integrated graphics processor (IGP); and (4) A high-speed link to the PCH chipset. The PCH's primary function is then to provide second tier I/O to devices such as additional PCIe lanes, 6Gbs SATA, M.2 NVMe ports for SSDs, high definition audio, USB ports, and Ethernet.



An Z170 chipset block diagram showing the connection of the i7-6700K CPU to the PCH, RAM, PCIe lanes, and other peripherals

Historically, chipsets have been the most disappointing aspect of any CPU release. For the longest time, Intel had a bad habit of providing somewhat brain damaged chipsets that did not meet the relative power of the CPU it supported and, by extension, the needs of the enthusiast buying the system. Typically these chipsets lacked enough latest-gen features such as enough USB 3.0 ports, SATA 6Gbps ports, or limited how its PCIe lanes could be provisioned to support multiple graphics cards.

Because the CPU in conjunction with the chipset drives the features exposed on the motherboard in terms of expansion capabilities, your choice of CPU is really not just about raw processor speed. Rather, it is critical to analyze the platform and what its expandability capabilities bring to the table to suit your particular purpose. The platforms are what differentiates the builds we will create later in the handout.

If, for example, you are a Viz Wiz using the latest in GPU-compute processing for renderings and animations, you may necessarily need more than one high end graphics card in the system. Then you necessarily require the additional PCIe lanes and slots that an HEDT / workstation class platform such as found in Broadwell E and Xeon. We'll discuss PCIe in depth in our discussion on graphics later.

Intel's 6th Generation Desktop Processor Lineup

Introducing Skylake – Intel's 6th Generation Tock

Skylake the code name for Intel's 6th generation microarchitecture using the same 14nm process as Broadwell. Recall that Broadwell was a "tick" – a die shrink of the preceding Haswell microarchitecture, but launched months late in mid-2015. Skylake was a new microarchitecture and already on tap for production, so Intel shortened Broadwell's release cycle to just 60 days and was never put into circulation as a mainstream desktop CPU. Instead, Broadwell was initially to mostly mobile platforms (where a die-shrink makes the most gains in increasing PPW). Now Broadwell is the latest standard for HEDT CPUs and the workstation-class Xeons, and Skylake for mainstream desktop. With Skylake's launch in August 2015, Intel hoped to recover from stumbling in their tick-tock cadence and get them back on track for the next round of new 10nm hotness. It did not exactly work.

Overall, Skylake averages about a 5-7% performance jump over Haswell processors of the same core frequency. That's perhaps less than the performance increases in microarchitecture advancements in the past, but this may be a case of "the new normal." Essentially, this really indicates that Intel is having difficulty having its internal architectural optimizations bear fruit as serious performance improvements over the previous generation. Of course, that is not to say that such optimizations are not worthwhile, just that they probably won't deliver any jaw-dropping surprises.

Skylake's Microarchitecture Improvements

Skylake's internal improvements over Haswell are differentiated in several ways. First, Skylake is intended to carry forth advancements across the entire power spectrum, from chips designed for tablet and mobile platforms through to the desktop, High End Desktop, and full-tilt workstation class CPUs. For now we shall concentrate on the things that are in the mainstream desktop lineup.

Major changes for motherboard manufacturers include the new LGA 1151 socket and the elimination of the fully integrated voltage regulator (FIVR) that was introduced in Haswell only two years prior. This reduces chip complexity and overall power draw, and allows motherboard manufacturers to provide their own voltage regulation implementations on the motherboard.

Skylake brings full DDR4 memory support to the desktop; previously, DDR4 was only available in HEDT and Workstation class systems. Because not everyone was not ready for DDR4 when it was introduced, Skylake also has backward compatibility with ***low-voltage (1.2V)*** DDR3***L*** memory. Note: *Using normal 1.5V DDR3 in a Skylake system can actually damage the CPU.*

Perhaps more importantly for AEC users, Skylake increases the maximum installed system memory to 64GB, up from 32GB in Haswell. Now that 16GB modules are commonplace, 64GB systems are easy to build with only 4 slots.

Internally, Skylake improves its ability to handle out-of-order instructions much better than Haswell, being able to process commands and data in fewer cycles and by handling/processing more data at once. Instead of waiting around for code to finish before it processes dependent code, the out-of-order architecture in Skylake puts many different branches of code and instructions in flight at once, where they can be grouped up in the scheduler and a level of instruction parallelism can be achieved.

Skylake also improves Hyperthreading performance, which basically overstuffs the execution pipeline in the CPU to where it presents two logical cores to the OS where there is only one physical core. Put simply, improvements related to HT allows each thread in a core to perform more efficiently, and its single threaded to multi-threaded performance ratio is better than in previous generations. which should mean that Hyperthreading in Skylake yields some tangible benefits when it comes to intensive workloads.

Speed Shift

Up until Skylake, power management was a task that was split between the CPU and the operating system. When the CPU is operating at the guaranteed rate – i.e. the core base frequency given in the spec sheets - Turbo Boost is enabled and performance ramps up and down via the CPU's power management hardware. However, when operating at frequencies below the guaranteed rate, such as when the system is not doing much computing or put to sleep, power management is done by the operating system. The system's firmware tells the OS a range of frequencies to choose from, and the OS picks one based on current workload, power settings, and system temperature. The issue is that power-down and power-up cycling could take a long time, reducing efficiency.

With Skylake, power management is more cooperative, with a new power saving feature called Speed Shift technology. The OS still has some say in the matter – the local power setting can force a lowered frequency to save battery life – but the CPU handles the rest. This makes power management far more responsive, allowing the system to "wake up" from a low frequency to high frequency in response to new workloads almost immediately, without waiting for the OS to give the go ahead.

Onboard Graphics are Still a Non-Starter

Skylake's on-board GPU, the Intel HD Graphics 530, is a rather mundane affair with still too-low performance to be worthy of consideration for AEC users. In fact, it underperforms the IGP on Broadwell by quite a margin. However, it is fully DirectX 12 compatible ensuring it will work with Windows 10 gaming titles which leverage it and beyond. While it drops support for VGA output, the HD 530 can, by itself, to run up to three 4K resolution (4096 x2304) screens using HDMI 1.4, DisplayPort, and DVI interfaces.

The Return of TSX Instructions

2014's Haswell microarchitecture *originally* was designed with a new extension to the x86 instruction set called **Transactional Synchronization Extensions New Instructions**, or **TSX-NI**. TSX-NI adds transactional memory support, which speeds up the execution of multi-threaded software. It monitors threads for conflicting memory accesses, aborting or rolling back transactions that cannot be successfully completed.

This is important for developers of multi-threaded applications (such as Autodesk) because writing stable, fast multi-threaded code is hard work. Programmers have to lock areas of memory in order to ensure concurrent threads do not overwrite the same memory location. They do this with complicated programming techniques that lock a wide region of memory (called a *table*) at a time. Called coarse-grain locking, this practice tends to cause delays and introduces more overhead to multi-threaded code execution and can stall out pipelines.

Consider two people trying to edit the same Excel spreadsheet at the same time. Even though they may be working on different parts of it, a coarse-grained approach is to lock the whole thing down while one users works on it until they finish, during which time the other user is left with nothing to do.

TSX instructions allows programmers to easily write fine-grain locks that lock smaller regions of memory and allow multiple threaded applications to perform much better. In our Excel example, this technique would only lock down individual rows or columns in the spreadsheet. This could potentially increase performance and efficiency, but with a higher risk of error. TSX-NI-able software would eliminate those errors by moving the evaluation of data integrity from software (which is slow), and into fast hardware. Initial tests indicate TSX-NI-enabled software will perform about 40% faster in specific workloads and provide 4-5 times more database transactions per second.

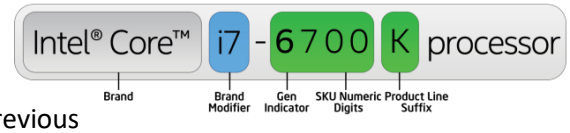
TSX-NI was intended to be made available in certain Haswell CPUs back in 2014. Unfortunately, in August of that year Intel confirmed a serious erratum (bug) found with the TSX-NI instruction set that could result in unpredictable system behavior, forcing Intel to disable TSX-NI **entirely** in all Haswell CPUs. However, the bug has been fixed and today, TSX-NI is fully enabled in Skylake.

To be of any benefit, Autodesk would have to code in TSX-NI support, and how this would affect overall performance remains to be seen. On the other hand, because there is no fix for certain CPUs, and AMD CPUs do not support them, Autodesk may opt not to develop any TSX-NI technology in its applications.

DMI 3.0 and Thunderbolt 3.0

Lastly, Skylake increases the I/O throughput to the chipset, up from 6GT/s in Haswell to 8GT/s over the new Direct Media Interface (DMI) 3.0 specification. Skylake also includes native support for Thunderbolt 3.0, which is a newer interface standard which combines PCI Express (PCIe), DisplayPort (DP), and provides DC power, all in one cable. Up to six peripherals may be supported by one connector through various topologies. In particular, Thunderbolt 3.0 uses the new USB Type-C connector, the "port of the future" which should be emerging as the new standard for ***all*** interfaces on modern PCs and mobile platforms.

How to Identify Intel Desktop CPUs



Before we get into the specifics, let us first review how Intel differentiates, classifies, and names their desktop 6th generation processors. While these conventions are specific to Intel's 6th gen CPUs, some are the same as previous generations. Make sure you consult Intel's documentation on their CPUs to verify all specifications.

Starting with what Intel terms the *brand modifier* (i.e., the prefix), the i7 series denotes a 4-core CPU with Hyperthreading (8 threads of execution). The i5 series is a 4-core without Hyperthreading (4 threads), and the i3 is a dual-core design with Hyperthreading (4 threads). Both i5 and i7 CPUs feature Turbo Boost; the i3 does not. The i3 also leaves out other extended Intel technologies, such as TSX instructions and vPro Technology.[17] It also comes with a lower L3 cache (3-4MB) vs. 6MB in the i5 and 8MB in the i7.

Skylake CPUs use the i7-**6**xxx / i5-**6**xxx /i3-**6**xxx name to indicate 6th generation processors; the other three numbers are stock-keeping unit (SKU) digits, or, more simply, model numbers.

The suffixes are where things get a little weird. The 'K' suffix designates a CPU with an unlocked multiplier, allowing the processor to be overclocked by the end user by tweaking the core clock multiplier in the BIOS[18]. K chips also sport the fastest core clock speed and the highest Thermal Design Envelope (TDP) which indicates how much power it uses in Watts. The i7-6700K has a TDP of 91W.

[17] http://ark.intel.com/compare/93366,90729,90733,90731,88195,37147,65523,63698
[18] http://www.velocitymicro.com/blog/intel-processors-locked-vs-unlocked-processor/

Other model suffixes available are: No suffix indicates that the core is not overclockable, and has a lower processor base frequency (3.4Ghz) and a lower TDP rating of 65W. The S suffix is for low-power models with a 65W TDP. The R suffix indicates a CPU meant for all-in-one, small form-factor, and other types of highly-integrated PCs and not available as discrete desktop chips. The T suffix stands for something Intel calls – and I'm not making this up – "power optimized lifestyle" models. These are ultra-low power models with a 2.8Ghz base frequency and only a 35W TDP.

Skylake i7-6700K Specifications

Comparing all of the 6th generation models, We can remove the S, R, T, as well as the entire i5 and i3 series from consideration. This leaves us with only one Skylake CPU model to choose from: the **Core i7-6700K**.[19] The i7-6700K was introduced back in 'Q3 2015, but with Intel's new "process–architecture–optimization" model, it is sticking around as the desktop CPU choice for perhaps longer than anticipated by the public.

Specifications are as follows:

Intel i7-6700K Specifications	
Number of Cores	4
Number of Threads	8
Processor Base Frequency	4.0 GHz
Max Turbo Frequency	4.2 GHz
Cache	8 MB
Thermal Design Power (TDP)	91 W
Max. Installed Memory Size	64 GB
Memory Types	DDR4-1866/2133 or DDR3L-1333/1600 @ 1.35V
# of Memory Channels	2
Processor Graphics IGP	Intel HD Graphics 530
PCI Express 3.0 Lanes	16
Pricing	Box : $339.00 / Tray: $350.00

Skylake i7-6700K Notes

Compared to Haswell from 2014, Skylake is much the same with improvements here and there. As we discuss in the section on graphics, the Z170 chipset provides an additional 16 PCIe 3.0 lanes, which may allow for two-GPU systems.

[19] http://ark.intel.com/compare/88196,88200,88195,93339

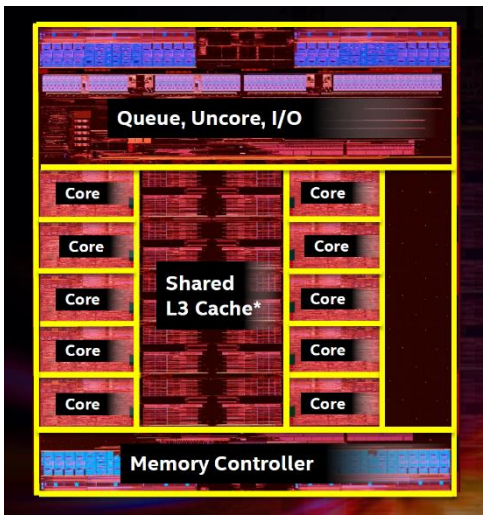
Intel's High End Desktop Platform with Broadwell E

The latest iteration of Intel's HEDT platform uses the new Broadwell E lineup of CPUs, which is a 14nm die shrink of the Haswell microarchitecture. Although a generation behind Skylake but providing more than four cores, Broadwell E is probably the best all-around high-end desktop-centric CPU available today for those that make full use of multithreaded applications. Primary specifications are below:

Intel Broadwell E Lineup				
Processor Name	**i7-6800K**	**i7-6850K**	**i7-6900K**	**i7-6950X**
# of Cores	6		8	10
# of Threads	12		16	20
Processor Base Frequency	3.40 GHz	3.60 GHz	3.20 GHz	3.00 GHz
Max Turbo Frequency	3.60 GHz	3.80 GHz	3.70 GHz	3.50 GHz
Cache	15 MB		20 MB	25 MB
Turbo Boost Max 3.0 Freq.	3.80 GHz	4.00 GHz	4.00 GHz	4.00 GHz
TDP	140 W			
Pricing (Tray | Box)	$434.00 | $441.00	$617.00 | $628.00	$1089.00 | $1109.00	$1723.00 | $1743.00
Max Memory Size	128 GB			
Memory Types	DDR4 2400/2133			
Max # Memory Channels	4			
Max # PCI Express 3.0 Lanes	28	40		

Broadwell E Model Notes

Broadwell E brings 14nm to the HEDT platform, which nets us 2 more cores in the 10-core i7-6950X in the same 140W TDP envelope of the 8-core 22nm Haswell E in 2015. Much is the same - we still get a mix of 28 to 40 PCIe 3.0 lanes and quad-channel DDR-4 memory. Broadwell E still uses the same X99 chipset and LGA 2011-3 socket as under Haswell E, which allows them to be a simple drop-in upgrade for Haswell E systems. The maximum installed memory size bumps up to 128MB, easily doable with X99's 8 DIMM slots and 16GB DIMM modules.



One thing that is different is price – Broadwell E systems are more expensive than Haswell E, sometimes by a wide margin. This is mostly due to the additional cores and higher clocks in Broadwell E.

Notice that the top of the line 10-core i7-6950X is only running with a base clock of 3 GHz, which is 25% slower than the 4GHz Skylake i7-6700K. This is directly due to the TDP limitation; to get all 10 cores firing away, any faster clock speeds would produce too much heat, so the highest base clock that would provide lowest TDP was chosen.

But it's **FAST**. At stock speeds and in benchmarks that stress multithreaded applications such as rendering, the Broadwell E i7-6950X is almost twice as fast as the Skylake i7-6700K, despite the 1GHz deficiency in clock speed. In benchmark tests there is between a 35 to 50 percent improvement going to the 6/8-core Broadwell E part over its simpler 4-core desktop cousin. Accordingly, performance with single-threaded benchmarks reveal the core clock deficiency. Still, 3.0GHz is still plenty fast for most operations.

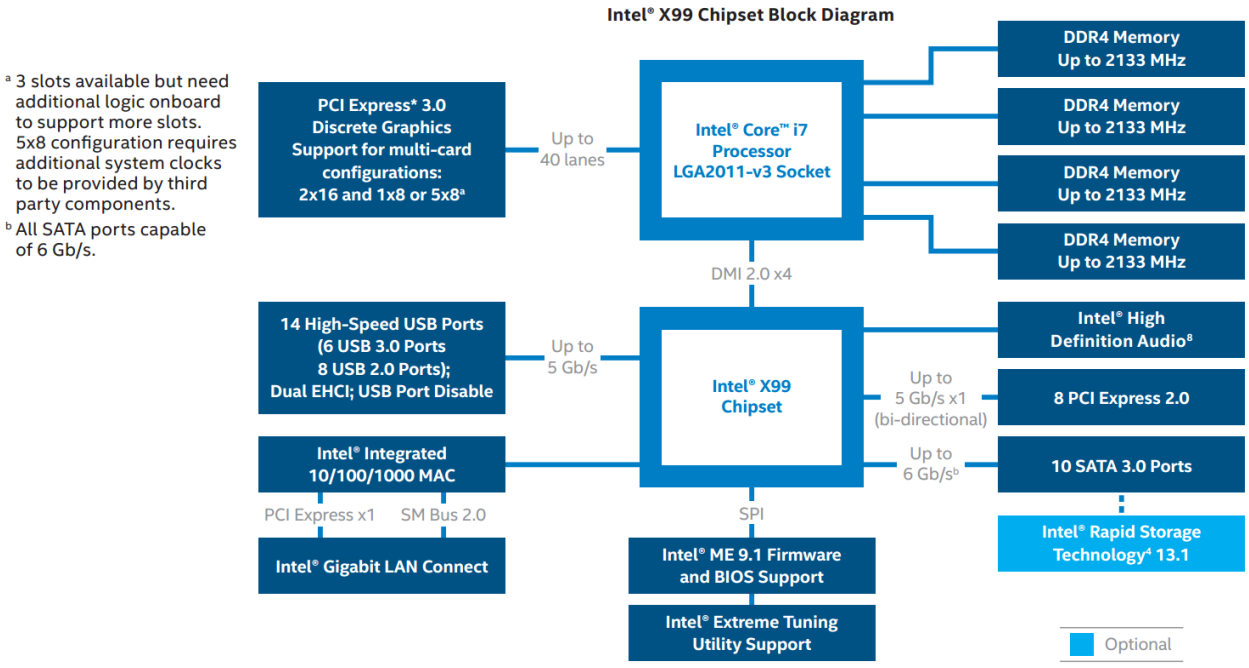
Another primary difference between Broadwell E and Skylake is the absence of an integrated graphics processor. The assumption here is that anyone opting for big desktop iron is going to outfit the system with at least one discrete graphics card - and possibly more - so the IGP is really quite useless. That is also one primary reason why Broadwell E includes more than 16 PCIe 3.0 lanes.

Speaking of PCIe 3.0 lanes, notice the i7-6800K's inclusion of only 28 PCIe 3.0 lanes to the others' 40 lanes. At first glance this may seem like a serious hindrance, but in fact this would only be limiting if you planned to stuff more than 3 GPUs into the system. With 28 lanes, dual GPU setups would run at x16/x8 speeds, rather than x16/x16 as you could with 40 lanes. Three-GPU setups on a 28-lane CPU would run at x8/x8/x8 instead of the x16/x16/x8 they would normally run with a full 40 lanes. However, given that any video card will run at PCIe 3.0 x8 speeds with very little to no degradation in throughput, this still provides solid GPU performance for low money.

Memory-wise, Broadwell E steps up speeds from DDR4-2133 to DDR4-2400. Any real-world performance gain to be had from increasing the memory speed this marginal amount is typically minimal for most operations. For those specific use cases that require fast memory (WinRAR compression, RAM Disks, in-memory virtual machines), there may sometimes see a benefit, but likely too little to matter.

One important note for system builders is that, for all HEDT processors, water cooling is almost considered must. You can possibly get by with a custom air cooling solution for the 6-core models, but it requires low-profile RAM due to the size of the heatsink required. Intel offers the same TS13X closed loop solution it did for Haswell E, and there are plenty of other all-in-one water cooling solutions on the market.

Broadwell E uses the LGA 2011-3 socket and is supported by the legacy X99 chipset. This modern PCH fixes some of the sins of the past by supporting up to six USB 3.0 ports and up to ten 6 Gb/s SATA ports. The weak link is that it only provides 8 PCIe 2.0 lanes, whereas Skylake's Z170 PCH provides an additional 16 PCIe 3.0 lanes. Being a legacy chipset, it also only supports DMI 2.0, with its 2GB/s interface to the PCH, instead of Skylake's new DMI 3.0, 4GB/s throughput. Gigabit networking and audio round out the package. New X99 motherboards support the new M.2 specification for PCIe-based SSDs.



Broadwell E and X99 Chipset Block Diagram

Xeon for Professional 3D Workstations

The Xeon line of CPUs specifically targets the professional 3D workstation and server market, and there is a broad range of cores, speeds, and cache to fit almost any need. Xeons are typically found in high end machines such as the Dell Precision and HP Z series of workstations.

However, understand that the term "workstation" is really just semantics. You can work just fine in all Autodesk AEC applications on a decent enough desktop machine, and play GTA V all day on a Dell Precision Xeon machine with 6 cores and gobs of RAM (I know this from personal experience). However, there are benefits to Xeon-based workstations for those that require its additional headroom and capabilities. The question to answer is this: Do you push the software in ways such that it would take advantage of a Xeon processor, or are you better served overall with more desktop CPUs like Skylake and Broadwell E? Combined with the choices you have based on preferred system vendor, and it's not an easy question.

For the DIY crowd, while you can build a Xeon-based system yourself, it will be a lot tougher. Xeon-capable motherboards are not as sold in the open market in the same volume as typical desktop components, and they have fewer differentiating features between them. You as a single buyer do not have the purchasing power of a Dell or HP, so they will most likely be more expensive than what you could get in a packaged workstation. You may find building a system based on the HEDT Broadwell E a better proposition.

Xeons traditionally live much longer shelf lives than their desktop cousins and there are many models out there, new and old. You need to review the CPU specifics in your workstation quote, as it may be outfitted with a Xeon from yesteryear, and you end up with older technology instead of the current generation.

Xeon - Key Technical Differences to Desktop CPUs

Xeons offers a few important internal differences over desktop CPUs:

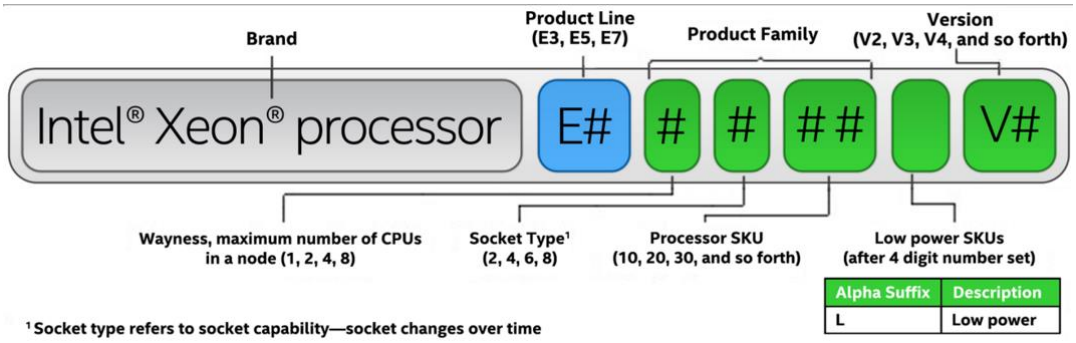
1. Xeons have better virtualization performance, with decreased time to needed to enter and exit a virtual machine. Xeons allow the shadowing of VM control structures, which should improve the efficiency of VM management. In addition, built-in hooks can monitor cache usage by thread so that "noisy neighbors" that affect cache contention can be isolated from other VMs.
2. Xeons are qualified to handle heavier, more intensive loads over a sustained time.
3. Xeons cannot be overclocked easily. Most Xeon-based motherboard are built for stability, not speed.
4. Only the lowest-end quad-core Xeons come with an integrated graphics processor (IGP).
5. Xeons have larger (sometimes much larger) L3 cache, e.g. 15-20MB vs. 8MB L3 cache on Skylake.
6. Higher-order Xeons can have 6 / 8/ 10 / 12 / 18 cores on a single die, but will run slower.
7. High-end Xeons support multi-CPU configurations. Desktop / Broadwell E are limited to one CPU.
8. Xeons support ECC (Error Correcting Code) memory. Skylake / Broadwell E do not.
9. Xeons generally support more installed system RAM (over 64GB) than desktop CPUs.
10. The Quick Path Interconnect (QPI) between CPUs or a CPU to the PCH is faster than even the DMI 3.0 interconnect used in Skylake and Broadwell E.
11. Xeon E3 models have dual-channel memory controller, but E5 Xeons, like Broadwell E, utilize quad-channel memory controllers, requiring RAM DIMMs to be installed in fours. However, this has little to no effect in real-world performance.[20]

[20] http://www.pcworld.com/article/2982965/components/quad-channel-ram-vs-dual-channel-ram-the-shocking-truth-about-their-performance.html?page=3

The Xeon Product Lines

Xeons are split into three major product lines. The E3 series is at the low end, basically Xeon versions of the 4-core desktop Skylake processors, and are meant for entry-level workstations. The E5 arguably has the broadest appeal, available in a wide variety of configurations, and is more akin to the Broadwell E / HEDT lineup of CPUs. You will find the Xeon E5 processor in single and multi-CPU, medium to high-end professional graphics workstations from name vendors. The E7 is in its own weird little world and is strictly for enterprise-level computing, such as database servers, Citrix servers, and in general powering the cloud. While the Facebook crowd isn't buying E7 based systems, Facebook itself has probably more than a few.

The naming convention for Xeons is markedly different from desktops but is highly informative at a glance. It follows the nomenclature described in this graphic:



The E# tells you the product code and in general what performance class you are in. The first digit in the 4-digit code after the dash tells you how many you can install in a system (1, 2, 4, and even 8). The second digit indicates the socket type it uses. The last two numbers are the specific product SKU, where the numbers ending in zero or one do not have an integrated graphics processor (IGP), and models ending in 5 or 6 have an IGP. "L" suffixed models are low power and thus are of no concern. The trailing version number (v3, v4, v5) is arguably one of the most important codes as it tells you what microarchitecture the processor is based on. Version-less parts are based on Sandy Bridge, v2 parts were based on Ivy Bridge, v3 parts are based on Haswell, v4 parts are Broadwell, and v5 parts are based on Skylake.

Xeons for Workstations

Xeons serve the high end workstation and in particular the virtual machine / database server markets well because they can handle many more VMs and database transactions without choking on the workload. But the demands on a BIM workstations are very different from that of heavy database and VM servers. For one thing, a single user is not going to peg the processor to the degree that a hundred thousand database transaction every second will. Much of your workload may be graphics related and not be entirely CPU-bound.

To that end, we are going to limit ourselves to a select subgroup of what is currently available. To make the cut for our needs, we are only going to review Xeons that meet the following criteria:

- We are only going to consider CPUs in the E3 and E5 product lines;
- We will consider only models based on the newest Skylake (v5) and Broadwell (v4) series;
- We will limit the maximum cores per processor to 8 and a minimum core clock speed of 3.2GHz;
- We will limit ourselves to models which support up to 2 physical processors;
- We will only consider models with no integrated graphics processor (IGP)

We limit ourselves to 2 physical CPUs / 8 cores max / 3.2Ghz because this represents the most practical (and affordable) BIM workstation configuration. While you can certainly stuff as many as 8 CPUs into a box with each CPU having upwards of 18(!) cores, the cost to do so is quite prohibitive and the return on almost any BIM related application is zero. Those sorts of configurations are more appropriate for VM servers who host many virtual machines at one time, or for rendering powerhouse machines. But as we will see later, building such massive boxes solely for rendering is not a worthwhile expenditure of money.

It is important to note that the more cores you have on a CPU, the lower the core clock speed must be to handle the thermal stresses. This will undoubtedly hurt single-threaded application performance that would otherwise scream on a desktop 4GHz i7-6700K. Much of what we do in BIM and 3D is still single threaded, where raw speed reigns over the number of cores. Most designers simply never spend enough time crunching rendering problems that would warrant more than 8 cores per CPU. This is particularly true today, now that we have very good cloud-based rendering services, in-house rendering farms, and GPU-based rendering solutions available at a very low cost that can provide imagery far faster than localized CPU-bound rendering solutions can.

Xeon E3-12xx v5 Family

For this class of Xeon, we are going to limit ourselves to considering only the latest E3-12xx v5 iteration, based on the Skylake microarchitecture. As stated previously, members of the Xeon E3 family are meant for entry level workstations and are really Xeon versions of corresponding desktop CPUs, with very similar specifications. The difference is that while there may only be one worthy desktop model (i.e., the i7-6700K), Intel builds four usable Xeons in this class, all similar but separated by 100 MHz clock speeds.

The E3-12xx v5 CPU works in single CPU only configurations and uses the FLGA 1151 socket. As with their desktop cousin, they come with 4 cores, all support Hyper-Threading, have a dual-channel memory controller supporting up to 64GB of both DDR4 and DDR3 RAM, 8MB of L3 cache, and support for the new TSX-NI instructions. Aside from the lower available clock speeds and support for ECC RAM, they are almost identical to the desktop Core i7-6700K. Given the minimum criteria listed above, the contenders are:

Processor Name	Cores / Threads	Base Frequency	Max Turbo Frequency	L3 Cache	TDP	PCI-E Lanes	Bus Type	Bulk Price
E3-1230 v5	4 / 8	3.4 GHz	3.8 GHz	8 MB	80 W	16	DMI 3	Tray: $250.00
E3-1240 v5	4 / 8	3.5 GHz	3.9 GHz	8 MB	80 W	16	DMI 3	Tray: $282.00
E3-1270 v5	4 / 8	3.6 GHz	4 GHz	8 MB	80 W	16	DMI 3	Tray: $339.00
E3-1280 v5	4 / 8	3.7 GHz	4 GHz	8 MB	80 W	16	DMI 3	Tray: $612.00

Source: http://ark.intel.com/compare/88182,88176,88174,88171

Xeon E3-12xx v5 Notes

1. All models end in a 0 (zero), which indicates they do not have an IGP.
2. Pricing is given in 1000-bulk trays for system integrators. Box retail prices are usually $11 more.
3. Any 100MHz increase in speed is at best a paltry 2-3% increase in performance, which is negligible.
4. Note the ***very*** high price jump of $273 from the E3-1270v3 @ 3.6 GHz to the E3-1280v3 at 3.7 GHz. Paying that much extra for a tiny 100MHz speed increase is not recommended.
5. Compare these specifications and prices to that of the Skylake i7-6700K running at 4GHz.

Xeon E3-12xx v5 Analysis and Conclusion

For the E3 family, I think the **E3-1270 v5** @ 3.6 GHz at $339 holds the sweet spot right now, but even then it's not all that attractive. Remember that the latest Skylake i7-6700K is running at a ***guaranteed*** *core clock* speed of 4 GHz (even when all four cores are active) for the exact same price. The E3 CPUs only reach 4GHz using Turbo Boost whose performance increase, as noted previously, depends entirely on the "Turbo Bins" for that particular CPU. Given the low range between minimum and maximum clock frequency, one can safely assume that 4GHz is reachable only when the CPU is handling single-threaded loads with one core active.

Given that the other specifications are almost exactly the same between the two, you should not pay a premium for a "workstation" class machine that is, internally, physically the same but slightly slower that a mainstream desktop. Clearly, there's no performance advantage to buying a Xeon E3-12xx workstation over a more mundane desktop i7-6700K.

However, you may be limited to CPU choice by your preferred system vendor. We look at this topic in more detail in a later section. In some cases your vendor may provide the i7-6700K as a cross-grade from a Xeon E3 model, as their specifications are very close and nothing really differentiates them.

Xeon E5-16xx v4 Family

The Xeon E5 v4 family represents the broadest implementation of Intel's Broadwell-based workstation / server class CPU and, with their increased number of cores, is generally recommended for BIM workstations that have to do a lot of multithreaded processing such as rendering.

As noted previously, E5-1xxx models are uniprocessor, meaning they go in workstations that fit only one physical CPU. The Xeon E5-16xx v4 is built on the Broadwell 14nm microarchitecture and very similar to the Broadwell E discussed previously. For the E5-16xx v4 we have five choices that are appropriate.

Processor Name	Cores / Threads	Base Frequency	Max Turbo Frequency	L3 Cache	TDP	PCI-E Lanes	Bus Type	Bulk Price
E5-1620 v4	4 / 8	3.5 GHz	3.8 GHz	10 MB	140 W	40	5 GT/s DMI 2.0	Tray: $294.00
E5-1630 v4	4 / 8	3.7 GHz	4.0 GHz	10 MB	140 W	40	5 GT/s DMI 2.0	Tray: $406.00
E5-1650 v4	6 / 12	3.6 GHz	4.0 GHz	15 MB	140 W	40	5 GT/s DMI 2.0	Tray: $617.00
E5-1660 v4	8 / 16	3.2 GHz	3.8 GHz	20 MB	140 W	40	5 GT/s DMI 2.0	Tray: $1,113.00
E5-1680 v4	8 / 16	3.4 GHz	4.0 GHz	20 MB	140 W	40	5 GT/s DMI 2.0	Tray: $1,723.00

Source: http://ark.intel.com/compare/82763,82764,82765,82766,82767

Xeon E5-16xx v3 Notes

1. The Xeon E5-16xx v4 family very closely resembles select Broadwell E CPUs in almost every respect. For example, spec for spec the Xeon E5-1650 v4 almost exactly matches the specs for the Broadwell E i7-6850K;[21] and the 8-core Xeon E5-1660 v4 closely matches specs with the Broadwell E i7-6900K[22]
2. All Xeon E5-16xx v4 CPUs connect to the chipset PCH via a 5 GT/s DMI 2.0 link (unlike the 8GT/s DMI 3.0 link in the Xeon E3-12xx v5 series), provide a 40-lane PCIe 3.0 controller, have a 4-channel memory controller and support up to 1.54TB of DDR4 RAM, with speeds up to DDR4 2400.

[21] http://ark.intel.com/compare/92994,94188
[22] http://ark.intel.com/compare/92985,94196

Xeon E5-16xx v3 Analysis and Conclusion

Comparisons to the high-end desktop Broadwell E are inevitable, as the **Xeon E5-1650 v4** is almost spec for spec identical to the Broadwell E i7-6850K – exactly the same number of cores, base clock speed, L3 cache, TDP, PCIe lanes, and even the same price. ***This 6-core Xeon is my pick of the litter in this class.***

At $1,113, the addition of 2 more cores in the 8-core E5-1660 v4 is not likely worth the $500 upcharge.

At $1,723, the 8-core E5-1680 v3 @ 3.4GHz is $610 more than the 8-core E5-1660 v4 for a paltry 200MHz increase in core clock speed. It's almost $700 more than the 8-core Broadwell E i7-6900K running @ 3.2GHz, and identical in price to the Broadwell E 10-core i7-6950X. While the i7-6950X, at 3GHz, is slower than our 3.2GHz minimum cutoff for consideration, reports show that it can be (and often is) overclocked to around 4.1GHz on average without much trouble. Bottom line: If you need 8 cores, build a Broadwell E i7-6900K system and forego the $700 Xeon tax.

Xeon E5-26xx v4

The Xeon E5-26xx v4 is also a Broadwell-based CPU but can be implemented on a single or dual-CPU system. While there many models to choose from, only three contestants meet our minimum specs:

Processor Name	Cores / Threads	Base Frequency	Max Turbo Frequency	L3 Cache	TDP	PCI-E Lanes	Bus Type	Bulk Price
E5-2637 v4	4 / 8	3.5 GHz	3.7 GHz	15 MB	135 W	40	QPI	Tray: $996.00
E5-2643 v4	6 / 12	3.4 GHz	3.7 GHz	20 MB	135 W	40	QPI	Tray: $1,552.00
E5-2667 v4	8 / 16	3.2 GHz	3.6 GHz	25 MB	135 W	40	QPI	Tray: $2,057.00

Source: http://ark.intel.com/compare/92983,92989,92979

Xeon E5-26xx v4 Analysis and Conclusion

Clock speeds for the E5-26xx series are a little slower than E5-16xx, between 3.2-3.5GHz, and their prices are twice as high on average. The strength of the lineup is the ability to work in dual-CPU configurations, so they should only be considered for such systems. Two 8-core E5-2667s with 16 cores / 32 threads will smack down everyone else when it comes to CPU-bound rendering tasks, but the slower 3.2 GHz base frequency means it will suffer – albeit slightly - in just about everything else. Additionally, such as system will cost over $4,100 for the CPUs alone. Consider it may be better to simply purchase another entire machine that could added to your local rendering farm.

Choosing a Mainstream Mobile CPU and Platform

When talking about choosing a mobile platform, you have many more choices. Ultrabooks, mainstream laptops, tablets, two-in-one tablet/laptop combos, and mobile workstations comprise today's form factors and overall power platforms.

Generally, your first task is to consider the platform first, then choose the CPU specifics within that platform. You need to identify how the machine will function in your life. Will it be a mobile ancillary machine to your desktop, where you do light duty modeling, take it on the road, or into meetings to show off models, and make presentations? Or will be it a truly mobile workstation and desktop replacement that needs to crunch through large Revit and 3ds Max models all the time?

Most mobile users, myself included, tend to have their laptops/tablets stapled to their sides at all times. The ability to do massive amounts of modeling or visualization work is curtailed by the additional and necessary functionality of Internet access, mobility, and communications, so you may not care that it can't render a scene in under two hours, but you do care if you can't even open it on a cross-country flight.

Regardless of any other consideration, you will likely look for as much power in the most mobile friendly platform available. Balancing power and portability is a catch-22 situation for BIM / 3D visualization purposes, as faster laptops are traditionally bulkier than slower ones, so the faster the machine you have, the larger, heavier, noisier, and generally more unpleasant it is use on a daily basis.

In the case of the ancillary machine, you would probably base your purchasing decisions on form factor first. You will likely value lightness and thinness over everything else. But even so, does it have enough juice to run the AEC applications you need at least marginally well for the purpose of design review and presentation? Can it be called upon to display a complex Revit model in the application and be able to navigate through the views, sheets, etc. without excessive lag?

In the case of a true mobile workstation, it needs to be powerful enough to run all of your AEC software all of the time. It will not be as powerful as a respectable PC, even one based on a desktop CPU. Luckily, the yearly increases in Power per Watt means that all components are getting smaller, cooler, and more efficient while still maintaining or increasing performance.

With the latest 14nm generation combined with advancements in storage and graphics, those once-beefy mobile workstations are getting more mobile and more powerful. Intel's overall strategy is to build out a broad and scalable Skylake platform that provides a microprocessor for every form factor and type of device out there on the market. To that end we are going to look at three main mobile families that focus on the ultra-portable market, the mainstream desktop market, and the workstation replacement market.

7th Generation Kaby Lake Mobile Core i7 Processors

Intel's "process-architecture-optimization" strategy is being borne out in its mobile offerings. In 2016 Intel further refined Skylake with a new, 7th generation "Kaby Lake" mobile processor. However, the models provided so far are inadvisable for the AEC market, as they are designed specifically for Ultrabooks such as the Dell XPS 13.

Ultrabook is Intel's moniker for the new high-end, slim laptop platform that is built on Intel's low-power hardware platforms. Basic specifications for Ultrabooks call for them to be under 23mm in thickness for screens 14" and above, and under 18mm thick for 13.3" screen or smaller. They must be able to run for at least 5 hours or more on a single charge, and use SSDs for data storage. Think MacBook Air

The Ultrabook platform is designed to offer good performance while focusing on efficiency. An Ultrabook will be able to handle most of ordinary (read: non-AEC) everyday activities at ease, while not requiring a lot of energy. However, Ultrabooks are not suited for extensive power-hungry tasks that AEC users are known for, such as BIM, rendering, Photoshop, visualization with Lumion, gaming, editing videos, etc.

Perhaps the biggest update in Kaby Lake is the addition of a new media engine, which can decode the most popular Ultra HD video formats on-chip. While previous CPUs were powerful enough to do so in software, the effect on battery life was a problem, particularly in power-sensitive platforms such as the Ultrabook. By shifting decoding into hardware, Intel can improve battery life and reduce heat, and is estimating three times the battery life when decoding 4K video. Not that playing YouTube videos more efficiently means a lot for most AEC users, but looking forward, it is a solid architectural improvement.

At some point we probably will get a Kaby Lake desktop processor. There are rumors of a Kaby Lake desktop i7-7700K model coming in early '17 [23], and even a Kaby Lake-X (along with a Skylake-X) for the HEDT market. Intel's strategy is to use these optimization phases to improve the mobile market first, then move on to the desktop and HEDT platforms, and finally to the enterprise / server (read: Xeon) market.

[23] http://wccftech.com/intel-kaby-lake-desktop-lineup-leak/

In the end analysis, however, it is clear that Kaby Lake isn't quite a true upgrade from Skylake. It's still 14nm, and it's still the Skylake architecture. Essentially we're stuck in a "tock" and Kaby Lake is really more like Skylake+. While the innards may be more energy efficient and faster, it is the implementation of the CPU in terms of cores, clock speeds, and cache that is really a determining performance factor.

In this series, Intel provides two 7th generation Kaby lake processors for consideration: the Core i7-7500U, and the Core i7-7Y75. Specifications are as follows:

Processor Name	Cores / Threads	Base Frequency	Max Turbo Frequency	L3 Cache	TDP	IGP	Bulk Price
i7-7500U	2 / 4	2.7 GHz	3.5 GHz	4 MB	15 W	HD Graphics 620	$393.00
i7-7Y75	2 / 4	1.3 GHz	3.6 GHz	4 MB	4.5 W	HD Graphics 615	$393.00

Source: http://ark.intel.com/compare/95451,95441

I7-7xxx Notes, Analysis, and Conclusion

As you can see above, both Kaby Lake CPUs are dual core with Hyperthreading. In fact, that is what the "U" in the i7-7500U means – ultra-low power draw. The "Y" in the i7-7Y75 means *extremely* low power, which explains the tiny 4.5W TDP and really low 1.3GHz core base frequency. Note that it can ramp up to a respectable 3.6 GHz, but as with all Intel processors, remember that the turbo bins of a specific processor determine under what conditions performance will drastically improve.

Today's Kaby Lake mobile processors are meant to power Ultrabooks and smaller powerful mobile platforms at a bare minimum TDP. Of the two, the i7-7500U is clearly more appropriate for basic AEC duties that would require decent multithreaded performance. However, with only 2 cores and a small 4MB cache, obviously neither is a prudent CPU for a desktop replacement, but more appropriate for the Ultrabook market with very light BIM and mostly presentation / email / web duties.

6th Generation Skylake Mobile Core i7 Processors

The next series to review is the 6th generation Skylake lineup of mobile CPUs. The following seven models are the ones you will see offered in most baseline and higher-performance laptop models. The shaded models in the following chart were introduced in Q1'16 and are refinements over the previous iterations.

Processor Name	Launch Date	Cores / Threads	Base Frequency	Max Turbo Frequency	L3 Cache	TDP	eDRAM	IGP	Bulk Price
i7-6700HQ	Q3'15	4 / 8	2.6 GHz	3.5 GHz	6 MB	45 W	None	HD Graphics 530	$378.00
i7-6770HQ	Q1'16	4 / 8	2.6 GHz	3.5 GHz	6 MB	45 W	128 MB	HD Graphics 580	$434.00
i7-6820HK	Q3'15	4 / 8	2.7 GHz	3.6 GHz	8 MB	45 W	None	HD Graphics 530	$378.00
i7-6820HQ	Q3'15	4 / 8	2.7 GHz	3.6 GHz	8 MB	45 W	None	HD Graphics 530	$378.00
i7-6870HQ	Q1'16	4 / 8	2.7 GHz	3.6 GHz	8 MB	45 W	128 MB	HD Graphics 580	$434.00
i7-6970HQ	Q1'16	4 / 8	2.8 GHz	3.7 GHz	8 MB	45 W	128 MB	HD Graphics 580	$623.00
i7-6920HQ	Q3'15	4 / 8	2.9 GHz	3.8 GHz	8 MB	45 W	None	HD Graphics 530	$568.00

Source: http://ark.intel.com/compare/88967,93341,88969,88970,93340,93336,88972

Introducing eDRAM

First introduced in specific Haswell models, embedded DRAM (eDRAM) is 128 MB of Dynamic RAM placed directly on the CPU die. Essentially, eDRAM is a new Level 4 (L4) CPU cache and is shared dynamically between the IGP and CPU, which should improve overall performance. However, eDRAM is Dynamic RAM, (i.e., the same transistor technology as in normal system RAM) not the Static RAM (SRAM) used in traditional L1, L2, and L3 caches. Thus it is slower, but much smaller, meaning it can be quite large in comparison. Being on-die, the performance hit from L3 to L4 eDRAM is much less than going from L3 to

main system memory across the front-side bus. Today eDRAM is actually pretty popular, as it is used in many devices such as the PlayStation and Xbox One gaming consoles and the Apple iPhone.

I7-6xxxHn Notes, Analysis, and Conclusion

1. All models are quad-core, support a maximum of 64GB of dual-channel DDR4-2133 system RAM and, being Skylake based, sport an 8 GT/s DMI 3.0 connection to the PCH. Note: Beware of any i7-66xx or i7-65xx CPU, as they are dual-core only, even with the i7 moniker.
2. Both the HD Graphics 530 and HD Graphics 580 IGP supports 4K resolutions on an integrated flat panel, external HDMI 1.4, and DisplayPort. None provide VGA output, which was removed in Skylake. All IGPs support DirectX 12 and OpenGL 4.4.
3. The "HK" suffix means high performance graphics, core unlocked. "HQ" is high performance graphics, quad core. Yes, you can now overclock an HK-equipped laptop.
4. While they are about $60 more expensive, I suggest you opt or the latest iteration 6x70HQ models that came out early this year – the i7-6770HQ, i7-6870HQ, and i7-6970HQ. Performance improvements via eDRAM can help make up for any 100 MHz drop in core clock speed, which help to keep the system running cooler and provide better battery life.
5. As with most other CPU lineups, notice the very large ~$200 jump to go from 2.7 GHz to 2.8 or 2.9 GHz. That slight increase will never be noticed in real life and simply makes the system run hotter.
6. Between all four legacy models, all are the same price ($378) but the i7-6820HQ is likely the better choice. Unless you want to overclock your laptop, which is really just lunacy, and opt for the HK.
7. Note the range between the base frequency and the turbo frequency. For single threaded, low-end workloads, the CPU can crank up into the mid 3GHz range, competing with desktop CPUs. With anything more stressful, the additional cores get busy and the clock frequency drops back to the guaranteed sub-3 GHz speed.
8. The i7-6920HQ is the performance king in terms of raw clock speed, but thanks to the additional 128MB of eDRAM cache, the i7-6970HQ should evenly compete with it and is therefore one of the fastest notebook processors in 2016. However, at $623 it is almost $200 more than the slightly lower clocked **i7-6870HQ**, which is my pick for the best mainstream mobile processor value.

Xeon E3-15x5M v5 Mobile Processors

In Q3'15 Intel introduced two new mobile E5 Xeon processors based on the Skylake microarchitecture at 14nm. In Q1'16 they added three new models which added eDRAM, similar to the updated mainstream i7-6x70HQ mobile CPUs mentioned in the previous section. These new models (shaded below) essentially supplant the previous two iterations, but they are all listed here for comparison.

Processor Name	Launch Date	Cores / Threads	Base Frequency	Max Turbo Frequency	L3 Cache	TDP	eDRAM	IGP	Bulk Price
E3-1505M	Q3'15	4 / 8	2.8 GHz	3.7 GHz	8 MB	45 W	None	HD Graphics P530	$434.00
E3-1535M	Q3'15	4 / 8	2.9 GHz	3.8 GHz	6 MB	45 W	None	HD Graphics P530	$623.00
E3-1515M	Q1'16	4 / 8	2.8 GHz	3.7 GHz	8 MB	45 W	128 MB	HD Iris Pro Graphics P580	$489.00
E3-1545M	Q1'16	4 / 8	2.9 GHz	3.8 GHz	8 MB	45 W	128 MB	HD Iris Pro Graphics P580	$679.00
E3-1575M	Q1'16	4 / 8	3.0 GHz	3.9 GHz	8 MB	45 W	128 MB	HD Iris Pro Graphics P580	$1207.00

Xeon Mobile E3-15x5M v5 Model Notes

1. All new models include 128MB of eDRAM.
2. Note the slowest of the new refresh Xeons is 2.8 GHz. This is the middle of the road for the i7-6xxx Skylake models noted previously.
3. Unlike desktop Xeon, these mobile Xeons include an IGP. The Iris Pro Graphics P580 uses system memory instead of dedicated video memory, although it will use the 128MB of local on-die eDRAM as a cache. Iris Pro Graphics are considered by Intel to be for "professional" applications such as CAD/CAM/BIM/etc. and does have drivers that are certified by Autodesk for Revit and AutoCAD, but not 3ds Max.
4. All five mobile Xeons support the same Intel technologies such as vPro technology, VT-d, TSX-NI, etc.

Xeon Mobile E3-15x5M v5 Analysis, and Conclusion

Mobile Xeons are more expensive than their mainstream mobile brethren, but except for the E3-1575M v5 (which is way out there) not by a whole lot. Note that they are all clocked slightly higher than their mainstream counterparts, e.g. the E3-1515M at 2.8 GHz ($489) vs. the i7-6870HQ at 2.7 GHz ($439). This 100MHz uptick in clock speed comes relatively cheap, at least by Intel's pricing standards.

Thus, my recommendation is the **E3-1515M v5**. At 2.8 GHz it's perfect for mainstream mobile heavy duty AEC users. You probably would never see the higher 100 MHz in the E3-1545M which is $190 more. And forget about the E3-1575M; paying a $718 upcharge for 200 MHz is just plain nuts. And remember the higher the core clock speed the hotter it will run, meaning that active cooling will run more often and the louder the system will be.

A Word on Evaluating CPU Benchmarks

This handout does not provide exhaustive benchmarks for processors. That's primarily because it's been difficult to impossible to find a steady series of independent benchmark tests that are applied to the several models that we would be most interested in. Most enthusiast web sites benchmark a CPU's gaming prowess over everything else, usually overclocking the CPU in the process and in general providing apples to oranges comparisons for the AEC market. These sites do not pay much attention to the workstation market and typically do not benchmark Xeon models; the ones that do still, for some reason, tend to emphasize differences in gaming performance, so we're back to apples to oranges.

Probably the best known site for examining benchmarks across a wide variety of low to high end CPUs is the CPU benchmark comparison chart at www.cpubenchmark.net. However, understand what it is: an average of user-submitted benchmarks from Passmark's CPUmark benchmarking tool. This should be a valid measure of a CPU's performance, and I believe it can validly be used for comparison between processors.

CPUmark measures raw CPU performance which may or may not reflect real-world results. The CPUmark score is mostly made up of benchmark algorithms which A) execute almost exclusively on the CPU and B) Fully uses the all the CPUs cores that are available. In effect, the CPUmark benchmark is CPU bound. However many real world applications are not CPU bound – they are disk intensive, require a fast video display, rely on high memory bandwidth, and so on. The CPU test has a small dependence on the RAM speed, so at least for the faster CPUs, better RAM can make the CPU look slightly faster.

Also, many real world applications are not very well threaded and run on only one CPU core. For these applications you won't see double the performance from a doubling in the CPUmark score.

For relative comparison, cpubenchmark.com lists the following Passmark CPUmark score results for the most popular processors of yesteryear as well as those discussed in this handout:

Processor Model	Cores / Threads	CPU Speed (Base - Max Frequency)	Passmark CPUmark Score
Lynnfield Core i7-860 (c.2009)	4 / 8	2.8 – 3.5 GHz	5,083
Sandy Bridge Core i7-2600K (c.2010)	4 / 8	3.4 – 3.8 GHz	8,498
Ivy Bridge Core i7-3770K (c.2012)	4 / 8	3.5 – 3.9 GHz	9,555
Haswell "Devil's Canyon" Core i7-4790K	4 / 8	4.0 – 4.4 GHz	11,187
Skylake i7-6700K	4 / 8	4.0 – 4.2 GHz	11,014
Broadwell E Core i7-6800K	6 / 12	3.4 – 3.6 GHz	13,615
Broadwell E Core i7-6850K	6 / 12	3.6 – 3.8 GHz	14,318
Broadwell E Core i7-6900K	8 / 16	3.2 – 3.7 GHz	17,431
Broadwell E Core i7-6950X	10 / 20	3.0 – 3.5 GHz	20,049
Xeon E3-1230 v5	4 / 8	3.4 – 3.8 GHz	9,624
Xeon E3-1240 v5	4 / 8	3.5 – 3.9 GHz	10,288
Xeon E3-1270 v5	4 / 8	3.6 – 4.0 GHz	9,959
Xeon E3-1280 v5	4 / 8	3.7 – 4.0 GHz	10,502
Xeon E5-1620 v4	4 / 8	3.5 – 3.8 GHz	9,933
Xeon E5-1630 v4	4 / 8	3.7 – 4.0 GHz	10,322
Xeon E5-1650 v4	6 / 12	3.6 – 4.0 GHz	14,372
Xeon E5-1660 v4	8 / 16	3.2 – 3.8 GHz	15,588
Xeon E5-1680 v4	8 / 16	3.4 – 4.0 GHz	17,060
Xeon E5-2637 v4	4 / 8	3.5 – 3.7 GHz	9,665
Xeon E5-2643 v4	6 / 12	3.4 – 3.7 GHz	14,329
Xeon E5-2690 v4	14 / 28	2.6 – 3.5 GHz	22,843
Xeon E5-2697 v4	18 / 36	2.3 – 3.6 GHz	23,070
Xeon E5-2698 v4	20 / 40	2.2 – 3.6 GHz	24,615
Xeon E5-2679 v4	20 / 40	2.5 – 3.3 GHz	25,236

Sources: https://www.cpubenchmark.net/high_end_cpus.html

Passmark's *CPUmark* is a processor benchmark and thus will naturally favor more cores, but sometimes results don't completely add up. For example, the 8-core Broadwell E Core i7-6900K @ 17,431 scored faster than the 8-core Xeon E5-1680 v4 @ 17,060, even though the i7-6900K has a slower clock speed.

Note the Skylake i7-6700K score of 11,014 beats every other 4-core CPU except for the older Haswell i7-4790K. This could be due to the higher turbo boost speed and the ability to overclock of the Haswell chip.

Also note the 10-core Broadwell E i7-6950X score of 20,049 is surprisingly close to the 20-core Xeon E5-2698 v5 score of 24,615. The Xeon has double the cores but is only about 23% faster. Clearly the large 800 MHz core clock speed difference is the reason behind this.

Bear in mind that Xeons are integrated into workstation-class systems which are primarily concerned with stability over speed, so have very conservative BIOSes with no overclocking ability, unlike the more DIY systems built on desktop and HEDT platforms.

The Final Words on Processors

The latest Skylake desktop and Broadwell E high-end desktop processors perform **very** well, coming close to or downright beating the Xeons in several tests, particularly ones that stress multimedia and single-threaded performance. Whereas the high-end E5 Xeons will rule in highly multithreaded application scenarios, such as working on very large Revit models and rendering on dedicated 3ds Max workstations, particularly with the E5-26xx series in a dual CPU configuration.

For the Grunt:

For most mainstream applications, as well as the probably 90% of your BIM Grunts, the Skylake i7-6700K CPU will do nicely at an appropriately moderate price point. The **i7-6700K** is the next evolution beyond last year's already excellent Haswell i7-4970K, and about 10% faster per clock. It is technically and financially equivalent to its Xeon cousin the **E3-1270 v5**, but is clocked a little higher. Both provide excellent baseline performance for designers of all kinds.

For the BIM Champ:

For advanced Revit users who push the software with Dynamo and integrate it with various other applications, a faster machine is in order. For these folks I like the **Broadwell E 6-core i7-6850K** with all 40 PCIe lanes for multi-GPU setups. If the user does any serious rendering in traditional renderers, the **Broadwell E 8-core i7-6900K** is a great but expensive alternative. Both processors will push performance much higher than the Skylake. The i7-6900K in particular is almost as fast as the very pricey i7-6950X.

For the Viz Wiz

For high-end 3ds Max Design use, where your need for fast local rendering trumps almost everything else, the **10-core Broadwell E i7-6950X** is really unparalleled. At a base core clock speed of 3.0 GHz, it will still handle single-threaded applications with aplomb, but will crush everything save a 14+ core E5-269x that costs hundreds more. Outfit it with at least 64GB of RAM with a great video card or three, and watch the images fly out of the machine.

Beyond this, if you really need as much horsepower as you can throw at your design and visualization problems, look at systems with dual physical CPUs and/or very high core counts. Near the absolute top of the line is the **Xeon E5-2697 v4**, with 18 cores (36 threads) and 45MB of L3 cache, which rockets away with a Passmark CPUmark score of 23,070, more than twice as fast as the Skylake i7-6700K. At $2,702 per CPU, it's a tough one to swallow, until you need those renderings out the door by the end of business.

If you are thinking of buying dual-CPU systems to shave down rendering times, you should consider three alternatives: (a) Distributed rendering across a farm using Backburner; (b) Using GPU-based rendering with Iray, and / or (C) Using Autodesk's cloud-based rendering or rolling your own with Amazon AWS or similar cloud-compute services. Any and all of these solutions are intrinsically cheaper and potentially much more powerful than a very expensive single machine.

V. System Memory

Memory has always been the least exciting thing to worry about when specifying a new workstation, but not anymore. Back in olden times (2012 or so) you bought 8 or 16GB or so of standard, off-the-shelf cheap DDR3 DRAM and you were happy, dangit. Conventional wisdom today is to specify 16GB for a basic "Grunt" level BIM workstation that perhaps doesn't do a lot of rendering, 32GB for a BIM Champ machine, and 64GB+ for a high-end 3D Viz Wiz machine that is expected to take on plenty of rendering.

Surprisingly, for most common workloads, there is perhaps a case of diminishing returns after 8GB in terms of raw performance in any one app, although the economies of scale and capabilities of modern motherboards usually make 32GB or more a smarter overall buy, as explained later in this section. As with any other modern workstation component, purchasing RAM in a new system is not so simple as there are various memory chip speeds and capacities that factor into the equation.

Your New RAM Standard: DDR4 SDRAM DIMMs

Today's workstation memory standard for Skylake, Broadwell E, and Xeon E3 / E5 systems is a 288-pin Double Data Rate Type Four, Synchronous Dynamic Random Access Memory, Dual Inline Memory Module, or DDR4 SDRAM DIMM for short. The term "Double data rate" means the module transfers data twice per clock cycle, effectively doubling the data rate compared to old-school SDRAM.

Because the choice of CPU is always paramount, and the RAM used is dependent on the CPU, you really don't have much choice of what kind of memory (DDR3 or DDR4) you can use. Rather, your choices are in capacity and DIMM data rate.

DDR4 has been a long time in the making and market acceptance has been slow, primarily because only Broadwell E and newer Xeons required it. Since Skylake has come out on the desktop in late 2015, DDR4 has gained much more market share and prices have naturally dropped close to DDR3 levels.

There are several key differences between DDR3 and DDR4. First, DDR4 operates at a lower voltage at 1.2V than DDR3 at 1.5V. This equates to about 18W in power savings over DDR3 in a typical system with 4 DIMMS. Perhaps not much, until you factor in the vast number of server farms and data centers with large scale architectures with thousands of DDR4 modules, and it adds up.

The other key difference is bandwidth. DDR3 starts at 800 MT/s (mega-transactions per second) up to as high as 2133 MT/s. DDR4 picks up where DDR3 leaves off at 2133 MT/s and goes up from there, massively increasing memory bandwidth. Whether that makes a real-world difference remains to be seen.

Unfortunately this increase in speed results in an increase in ***latency*** as well. You will see this listed as CAS (Column Access Strobe) Latency, or CL, in memory specifications. It is the delay time between the moment a memory controller requests a particular memory column on a RAM module and the moment the data from the given location is available on the module's output pins. CL is given as a simple integer and the lower the CL, the better the module performs.

While a DDR3 module may have a CL of 9 or 10, DDR4 modules may have a CL of 13 or higher. However, given the higher bandwidth, the increase in latency is either a negligible or non-existent problem. If you are purchasing a packaged system, you will have no say in the specific modules you get, but if shopping for a DIY system and selecting individual components, you likely want DDR4 modules with the lowest CL value. Note that you need to ensure the CL number is the same for all DIMMs in the system.

The DDR4 DIMM module is physically but subtly different from DDR3 in several ways. First, they are pin-incompatible, meaning you cannot physically install the wrong module in a system. DDR4 are 288-pin DIMMs whereas DDR3 are 240 pin modules, and the pitch between the pins has been reduced from 1.0 mm to 0.85 mm in DDR4.

Physically there is a change in the bottom edge connector for DDR4, which is curved down in the middle area somewhat that lowers the insertion force required for installation in a motherboard slot.

DDR4 DRAM (left) and DDR3 DRAM (right). Image courtesy anandtech.com

Understanding DDR DRAM Data Rates, Speeds, and Naming Conventions

DDR4 DIMMs can be specified and purchased at different data throughput ratings, provided as two standard naming formats. DDR4 XXXX is the "friendly name" where XXXX is the transfer rate expressed in millions of data transfers per second, or Mega-Transactions/s (MT/s). You will also see the "module name" as PC4 NNNNN, where NNNNN is the total bandwidth of the module in MB/s.

Memory bandwidth is a measurement of the theoretical transfer rate of a communications channel, and is determined by a simple formula: **Bandwidth = Memory clock frequency x Bits transferred per clock cycle / 8** (as there are 8 bits per byte)**.** Since memory module DIMMs are 64-bit devices, the (memory clock * 64 bits per cycle) = the DDR transfer rate in MT/s. Thus, the entire formula is simplified down to **Bandwidth = DDR Transfer Rate x 8.** Thus, for any DDR4 XXXX module, its corresponding PC4 NNNN value is simply XXXX x 8, rounded down to the nearest 100.

Note that none of these numbers refers to the module's clock speed or frequency in MHz, GHz, or whatever. ***Many enthusiast sites, online stores, and even Crucial.com get this wrong all of the time***. In fact, the way the math works out, the data rate MT/s is in reality twice that of the I/O bus clock, due to the memory module being double-data rate RAM. For DDR4 running in a quad-channel memory controller (i.e., in HEDT or Xeon E5 systems), the I/O bus clock is then 4x that of the true memory clock, because DDR4 expands the memory bus to 256 bits, accessing 4 memory modules at a time (i.e., 4 x 64).

Some examples of current popular DDR module data rates in quad-channel mode are:

DIMM Type	Memory Clock	I/O Bus Clock	Data Rate (MT/s)	Module Name	Peak Transfer Rate
DDR4 2133	266 MHz	1066 MHz	2133 MT/s	PC4 17000	17 GB/s
DDR4 2400	300 MHz	1200 MHz	2400 MT/s	PC4 19200	19.2 GB/s
DDR4 2666	333 MHz	1333 MHz	2666 MT/s	PC4 21300	21.3 GB/s
DDR4 2800	350 MHz	1400 MHz	2800 MT/s	PC4 22400	22.4 GB/s
DDR4 3000	375 MHz	1500 MHz	3000 MT/s	PC4 24000	24 GB/s
DDR4 3200	400 MHz	1600 MHz	3200 MT/s	PC4 25600	25.6 GB/s
DDR4 3333	416.67 MHz	1666 MHz	3333 MT/s	PC4 26600	26.6 GB/s

However, understand that higher rated DDR4 RAM may not provide much if at all speed benefit by itself; it is very much dependent on the specific applications you use. Outside of database applications, RAM throughput doesn't scale well, so paying more for special very fast RAM is of no real benefit. Today's benchmarking indicates that DDR4 2666 is likely the sweet spot, with faster modules producing inconsequential gains. On the other hand, being a database application, Revit may respond better than most to increased DDR4 data throughput rates than those being benchmarked.

On-Die Memory Controllers, DDR RAM Channels, and Configuration Rules

When specifying memory for your system, it is more important to understand some of the internals of your particular CPU than consider faster RAM modules. Because CPU models have different memory controllers, you have to configure your memory correctly for peak performance.

Modern CPU architectures since Intel's Nehalem in 2007 (and AMD's Hammer in 2001) have embedded, on-die memory controllers which directly talks to the memory DIMM through two, three, or four separate channels. These channels effectively double, triple, or quadruple the communication bandwidth between the CPU and system memory.

The memory controller (read: the CPU model) ultimately decides what memory specifications it will support. Both Skylake and Xeon E3 CPUs have a dual-channel memory controller, supporting up to 64GB of DDR4-1866/2133 DIMMs. Broadwell E has a quad-channel memory controller, supporting up to 128GB of DDR4-2400/2133 DIMMS. Both the Xeon E5-16xx and E5-26xx have quad-channel controllers supporting up to 1.54TB of DDR4-1600/1866/2133/2400 DIMMs. Without modification, a DDR4-3200 module dropped in a Xeon E5-26xx will run at the top speed it supports, 2400 MT/s. That said, many motherboards can tweak their memory clocks in the BIOS to support higher-rate DDR4 modules.

Each channel is a 64-bit pathway to one (or more) memory modules. The controller talks to all of the DIMMs on these channels simultaneously and in parallel, and together are treated as a single entity from the CPU's point of view.[24] As the number of channels increases, so does the number of DIMMs talked to, and the overall bandwidth increases.

For a dual-channel memory controller, you need to install identical DIMMs in pairs so you can access both simultaneously. Similarly, for a quad-channel controller, you need install identical DIMMs in fours.

By talking to a full complement of DIMMs installed on the controller's number of channels, your CPU can minimize the effects of high latencies and timing delays that occur with fewer memory modules. Otherwise, RAM falls back to single data rate speeds and system performance suffers. Memory slots are color coded to indicate the channels, so you know to install RAM to fill the same-colored slots before moving on to fill others.

An X99 motherboard with 8 slots with a color coded pair of four channels each. Install 4 DIMMs in one colored set of quad-channels, then 4 DIMMS in the other colored quad-channels. Refer to the motherboard manual for memory DIMM compatibility and installation instructions.

The next constraint is the number of DIMM slots. You typically get two slots per channel, so dual-channel Skylake and Xeon E3 systems usually have four RAM slots. Broadwell E and Xeon E5 typically have 8 slots to provide a pair of quad channels. When installing sets of pairs or quads, while each DIMM inside a pair or quad must be identical, each pair/quad can have a different total capacity. For example, in a dual-channel system with four slots, you could install two 16GB modules and two 8GB modules, for 48GB total.

Thus, if you aren't maximizing your RAM capacity when you purchase the system, you want to determine the module size such that it completely fills up 4 slots (channels) with the largest capacity at a time, so that you don't populate all of your slots with smaller modules and leave it problematic to add more memory later.

[24] http://www.hardwaresecrets.com/everything-you-need-to-know-about-the-dual-triple-and-quad-channel-memory-architectures/1/

DDR Density and Pricing

DDR4 memory has the advantage that it can support single modules of higher densities, so 16GB, 32GB, on up to 512MB modules are theoretically possible. DDR3 has a theoretical maximum of 128GB per DIMM. Today, modules are available in 4, 8, and 16GB per module. Checking Newegg.com prices, 4GB DDR4 modules can be found for about $23, or $5.75/GB. 8GB DDR4 modules are around $40, or $5/GB, and 16GB modules are going for $100, or $6.25/GB. So the price differences per GB is really peanuts, when you consider everything.

The rule of thumb is to concentrate on the intended amount of RAM you think you need and purchase the least number of highest-capacity modules to get to that number, which can leave room for future upgrades. Next you have to consider the memory controller requirements and number of available slots, and optimize accordingly. Some common configurations are:

1. Skylake and Xeon E3-12xx systems support up to 64GB of RAM in two channels, so max out the available slots by installing 4x16GB modules. Use 2x16GB modules for a 32GB system. Using 4x8GB modules means you cannot upgrade to 64GB without starting over with 16GB DIMMs.
2. Broadwell E's quad-channel memory controller can address up to 128GB, and X99 motherboards have 8 slots, so you would need to install 8x16GB modules to max out the RAM. Common configurations are 4x16GB DIMMs for a 64GB system with room to grow. Likewise 4x8GB modules would get 32GB and still enable quad-channel bandwidth, with 4 slots open for an upgrade.
3. For high-end Xeon E5-26xx workstation-class systems that can address up to 1.54 TB (!) of RAM, memory configuration is also determined by the number of CPUs installed. Some systems have 16 slots, but are allocated to 8/CPU, so to use more than 8 slots you need a second CPU installed.

ECC Memory is not Required in a Workstation but RDIMMS May Be.

On Xeon-based high-end workstations, you will most likely find RAM in either ECC or Non-ECC flavors. ECC stands for Error Correcting Code, which can detect and correct most common kinds of internal data errors and corruption. ECC memory is specified in systems where data corruption cannot be tolerated under any circumstances, e.g. servers or scientific / financial computing environments.

Support for ECC memory is limited strictly to the Xeon and not supported on Skylake or Broadwell E platforms. ECC memory is more expensive than non-ECC and the error correction mechanism lowers memory performance by about 2-3 percent. Graphics workstation configurations by Dell / HP will often default to ECC memory, but in many cases you can specify non-ECC RAM without issue. Memory manufacturers have much tighter tolerances on their fabrication processes and the propensity for memory errors is much lower than it was just five years ago.

For these high-end systems with memory requirements over 64GB, you may need to specify **Registered DIMMs**, or **RDIMMs**. Most DIMMs in typical desktops are unregistered, unbuffered DIMM modules (sometimes called UDIMMs). For workstation and server-class machines, RDIMMs provide more capacity in that you can place more DIMMs per channel (DPC).[25] Typical systems allow up to two DPC, which is why you have 4 slots for dual channel and 8 slots for quad channel, limiting yourself to 2DPC x 4 channels x 16GB DIMM = 128GB. With RDIMMs you get up to (3DPC x 4 channels x 16GB DIMM = 192GB per CPU.

In very high end exotic systems you may need to go with **Load Reduced DIMMs**, or **LRDIMMs**[26] which can increase RAM capacity per CPU even more.

[25] http://www.anandtech.com/show/6068/lrdimms-rdimms-supermicros-latest-twin/2
[26] http://www.simmtester.com/page/news/showpubnews.asp?num=167

VI. Graphics Cards

Selecting the right graphics card is largely specifying the correct graphics processing unit, or GPU, for your AEC applications. This is often problematic, confusing, and downright irritating to say the least. There is a lot of fear, uncertainty, and doubt over how GPUs function in general and specifically how Autodesk applications use them. Once you understand some basics, it's actually not that hard, and this section will attempt to dispel some misconceptions.

To a large degree, the capabilities and performance of the graphics card is proportional to the cost of the card, but only to a point. Today, simpler operations such as 2D CAD are universally covered adequately across all cards that range from the low end to well over $1,000.

What separates the better graphics cards from the pack is how much time they can save you in wholly GPU-bound operations, such as orbiting in shaded 3D views, rendering using the Iray renderer, or pushing out complex animations in Lumion. These are the differentiating factors that go into a decision; if you do those things all of the time, you need to pick a strong performer in that area. Otherwise, you could perhaps pick a more moderately priced card. Or pick something perhaps up the scale that you can grow into as you adopt more GPU-bound operations in your daily work. The key point is that you need to let your computing needs determine the class of card you need.

Professional and Gaming Video Cards

The choice of graphics solution for any graphics and design workstation machine starts by looking at the two primary kinds of cards being offered. In one corner you have the expensive, "professional" workstation-class cards meant for the "true design professionals." In the other corner you have a plethora of cheap "gaming" cards meant for "dudes living in their parent's basements."

It may surprise you, but the dirty secret is that the GPU architecture used in professional cards like AMD's FirePro and Nvidia's Quadro lines are almost always identical to those used in their Radeon and GeForce line of gaming cards. The GPUs may have different code names but internally they are remarkably the same. For example, the architecture used in the top shelf AMD FirePro W family is the same design used in the Radeon HD 7xxx series desktop boards. On the Nvidia side, the $1,800 Quadro M5000 uses essentially the same GM204 GPU used in the $510 GeForce GTX 980. AMD and Nvidia may fine tune the innards of each for a particular target price point (typically by tweaking core clock rates), but the vast majority of the plumbing is the same. Carefully examining of benchmarks proves this out.

Even then, Nvidia and AMD still make a lot of money by selling professional level cards with professional level markups attached to them. There are several reasons why professional cards cost what they do:

1. Professional graphics cards have drivers that are certified by Autodesk, Dassault Systemes, Solidworks, and a dozen other application developers, to ensure the card meets the minimum specification under each application. This is expensive and costs Nvidia / AMD money. However, there is clear evidence to suggest that this can potentially means quite a bit, particularly for applications who are built around OpenGL instead of DirectX. Cards which are not certified may not be able to run those applications well or at all because their drivers simply aren't up to the task. On the other hand, certified drivers are not in and of themselves guaranteed to always work, either. I have seen plenty of "solid" card / driver combinations that are blessed by Autodesk fail in Revit or 3ds Max or both.
2. Professional-class drivers may take advantage of specific features in those cards meant for high-end content creation, such as enhanced antialiasing, z-buffering, and others. Gaming card drivers may or may not expose that functionality because those features are considered the domain of professional content developers, not consumers (gamers).

3. Workstation class cards may include more onboard RAM, although today this is somewhat moot as modern game titles (and gaming enthusiasts) demand more video memory. For Autodesk AEC users, the amount of onboard video RAM is most critically important for Iray rendering on Nvidia-based graphics cards.
4. AMD and Nvidia oversee the fabrication of their professional line of graphics cards with only a small number of manufacturers. This small number of vendors reduces competition which raises prices. Gaming cards are built by many manufacturers from reference designs from AMD and Nvidia.
5. Professional cards have tighter tolerances for build quality. These are generally built to run under more continuous duty conditions, like inside of server racks 24/7/365. To that end they are clocked slower to lower TDP and lengthen their lifespans. Gaming cards are meant for speed first and foremost, and aren't really built to run full tilt all day long. However, the demands of intense GPU based rendering can take its toll on the silicon, and causes temperatures to rise and cooling fans to work overtime. Gaming cards have been known to go up in smoke - literally - when pushed to render many scenes. But they are so inexpensive compared to professional cards, it's a wash.

While it's a given that professional cards can and do well in Autodesk applications, the strategy by Autodesk, workstation vendors, and Nvidia/AMD to promulgate professional-class cards only is hindered by two facts: First, the vast majority of Autodesk's AEC applications are largely graphics card agnostic. They have to be, in order for them to work across most PCs. All Autodesk applications use the traditional Windows DirectX 9 API from Windows 7 as a minimum standard, and all of today's graphics cards have passed that minimum standard long ago and should work fine.

That's not to say that a professional card won't work as well or better, but the idea that you have to run a professional card in any Autodesk application is pure nonsense. Both Nvidia and AMD understand how to code their drivers for the Windows environment very well. Professional cards have vetted drivers which may reduce the risk of seeing graphic anomalies in Revit, Navisworks, or 3ds Max, but nothing is guaranteed. Many people run el-cheapo gaming cards in high end Autodesk applications without issue.

Does Autodesk Certification Matter?

You may hear the argument that, because Autodesk has not certified a particular card, it should be discounted immediately as not appropriate. Remember that Autodesk is in publishing design software, not making sure your graphics card works. It has limited resources to test new hardware and the hundreds of drivers that appear almost daily, so it is likely difficult for Autodesk to keep on top of this. Checking the Autodesk certified hardware database[27], two things are apparent: First, nothing has been tested on 2017-era software, and the number of "certified" cards certified for Revit under Windows 10 is in fact very small; currently it is limited to the mobile Quadro M series.

Of course, there is no guarantee that you won't have issues with gaming cards as well. Many gaming cards and driver combinations work fine across the board for all Autodesk AEC applications. There are some I have found which don't play well with Revit or 3ds Max or both. This could be caused by a specific driver version, a particular manufacturer, model, card manufacture lot, host machine PCIe 3.0 timings, or combination of factors.

That said, in every case where I have seen driver-related issues in both professional and gaming cards, I've been able to find older or newer drivers that fix the issue. It may take some time to troubleshoot, but you can safely gamble that any time spent to fix a gaming card will be nowhere near as much as a professional card would cost. And if a particular gaming card simply doesn't work no matter what, sell it and get something else. At least you aren't overspending on something that is comparatively underpowered.

[27] http://usa.autodesk.com/adsk/servlet/syscert?siteID=123112&id=18844534&results=1&stype=graphic&product_group=2&release=2016&os=524288&manuf=1&opt=2

Professional vs Gaming Card Value

The second issue with professional cards is that in many cases, their relative cost/performance ratio gets seriously out of whack, and you end up paying a lot more money for the same or even lower performance that you would have with an unwashed gaming card.

For example: Last year's Nvidia GeForce GTX 970 card costs around $350, and by all accounts will work effectively well for all AEC application work, including Revit and 3ds Max Design. Based on the second generation Maxwell 2 architecture, it features the GM204 GPU with 1664 CUDA cores, a 256-bit memory interface, a 1050 MHz clock, and has 4GB of onboard RAM. It also has a dual-link DVI-I, HDMI 2.0, and 3 DisplayPort 1.2 ports, so it will plug into anything on your desk. By all accounts it represents a solid medium range value. Not the best for everything, but very good nonetheless.

On the other hand, an entry-level workstation class Nvidia Quadro K2200 is based on the 1st generation Maxwell architecture GPU GM107 with a paltry 640 CUDA cores, a 128-bit memory interface, a 1000 MHz clock, 4GB of RAM, 1 DVI, no HDMI, and only two DisplayPort 1.2 ports. It currently runs around $399 on up, retail. Although core clock rates are very close, it has 1/3 the CUDA cores, and its fill and texture rates are thus a fraction that of the GTX 970. Why would anyone pay possibly a lot more for something a generation behind that would provide absolutely no benefit?

In terms of performance, conventional wisdom states that just about any gaming card you buy priced from $250 on up will be fine for AEC application usage. To get the same performance from a workstation card could cost much more. Below that price point, you could suffer lag in dense point clouds, high polygon 3ds Max models, Revit views with shadows turned on, etc. Spend more to get more usable eye candy on screen, hardware-accelerated anti-aliasing, or if you simply need to push a lot of polygons around on a regular basis.

Introducing Nvidia's Pascal Architecture

Just as Intel and AMD have architectures for CPUs, so do graphics companies have architectures for GPUs. In 2016 Nvidia debuted the Pascal architecture, the successor to the highly successful Maxwell architecture that powered the GeForce GTX 9xx and M-series Quadro cards.

At a high level, Pascal is a mix of old and new, building on top of 2014's Maxwell more than anything else. Since the introduction of the Kepler architecture in 2012 and through the lifespan of the Maxwell architecture from 2014-2016, Nvidia has been effectively stuck at a 28nm fabrication process node. Although still at the same 28nm process, Maxwell was still able to make huge architectural advancements over Kepler, vaulting Nvidia forward in maximizing energy efficiency. Maxwell gave us the GeForce 700 and 900 series, as well as the M-series of Quadro graphics cards.

Pascal, on the other hand, is characterized by the wholesale move down to the 14nm/16nm FinFET process technology, enabling more features, higher performance, and improved Power per Watt. FinFET is a newer transistor technology that supplants the traditional High-K Metal Gate technology on older, larger processes. At 20nm, HKMG transistors were simply too leaky, and FinFETs were absolutely necessary to get good performance below 28nm.

At a 14nm/16nm FinFET process, Pascal provides the needed headroom to bring down power consumption and reduce chip size while pushing the envelope on GPU performance through all form factors, from ultrabooks and laptops through to graphics workstations and beyond. Pascal is characterized primarily as pouring on the clock speed to push total compute throughput to almost 9 Teraflops, and updating their memory subsystem to properly feed the GPU.

Nvidia Naming Schemes

Just like Intel, Nvidia has their own naming schemes for their GPUs based on architecture, generation, and performance. For example, let's take the GM204 GPU.

Gx xxx: The first letter is a constant which stands for Graphics (Technology)

xM xxx: The second letter is a variable which stands for the architecture of the chip. For Kepler it was K, for Maxwell it was M, for Pascal it is P, for Volta it will be V, and so on.

xx 2xx: This numeral stands for the generation of the architecture. First generation Maxwell, for example, had a 1 here. Similarly, GP100 is a first gen Pascal GPU.

xx x0x: This numerical has a variable meaning and can stand for everything from a better binned chip to a refresh. This digit is usually ignored when interpreting chips.

xx xx4: The last digit is the actual performance indicator of the chip and follows an inverse approach. Lower means higher performance, so a 0 here would mean the flagship die, and a 7 or 8 would mean the weakest die.

Pascal for HPC and Consumer

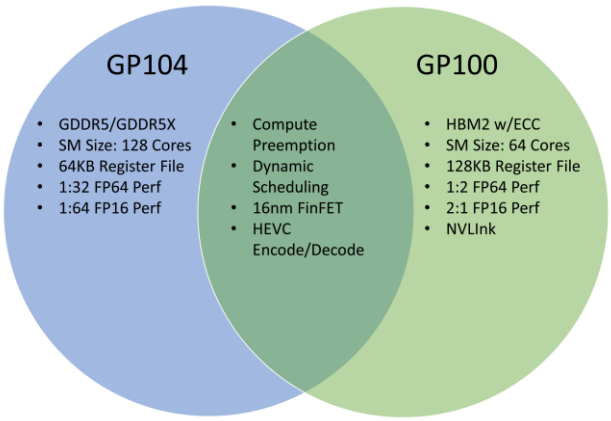
Pascal differs from Maxwell in that it diverges to cater to two markets simultaneously: the High Performance Computing (HPC) crowd and the consumer market. Although Nvidia had clearly separated GPU lines between HPC and consumer before in Fermi, Kepler, and Maxwell, with Pascal the difference in consumer and HPC GPUs is even more stark.

Pascal was originally first unveiled with the Nvidia's Tesla P100 powered by the GP100, a specialty GPU that is designed from the ground up specifically for HPC applications across many areas, including computational fluid dynamics (CFD), medical research, financial modeling, quantum chemistry, energy discovery, and several others. Tesla GPUs are installed in many of the world's top supercomputers and datacenters use Tesla GPUs to speed up numerous HPC and Big Data applications, as well as enabling Deep Learning and leading edge Artificial Intelligence systems. HPC and specifically the GP100 is at the forefront of Nvidia's work on self-driving cars.

Being an HPC GPU, the features specific to the Tesla P100 and GP100 are beyond the scope of this handout, but you can read about the developments in Nvidia's Tesla P100 whitepaper available at https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf

Pascal on the Desktop with the GP104

On the desktop GPU side, the GP104 powers Nvidia's flagship GTX 1080 and GTX 1070 graphics cards. The GP104 has quite a few features from the GP100 stripped away specifically to differentiate the two markets, although it keeps much of the core Pascal architecture that matters to consumers.

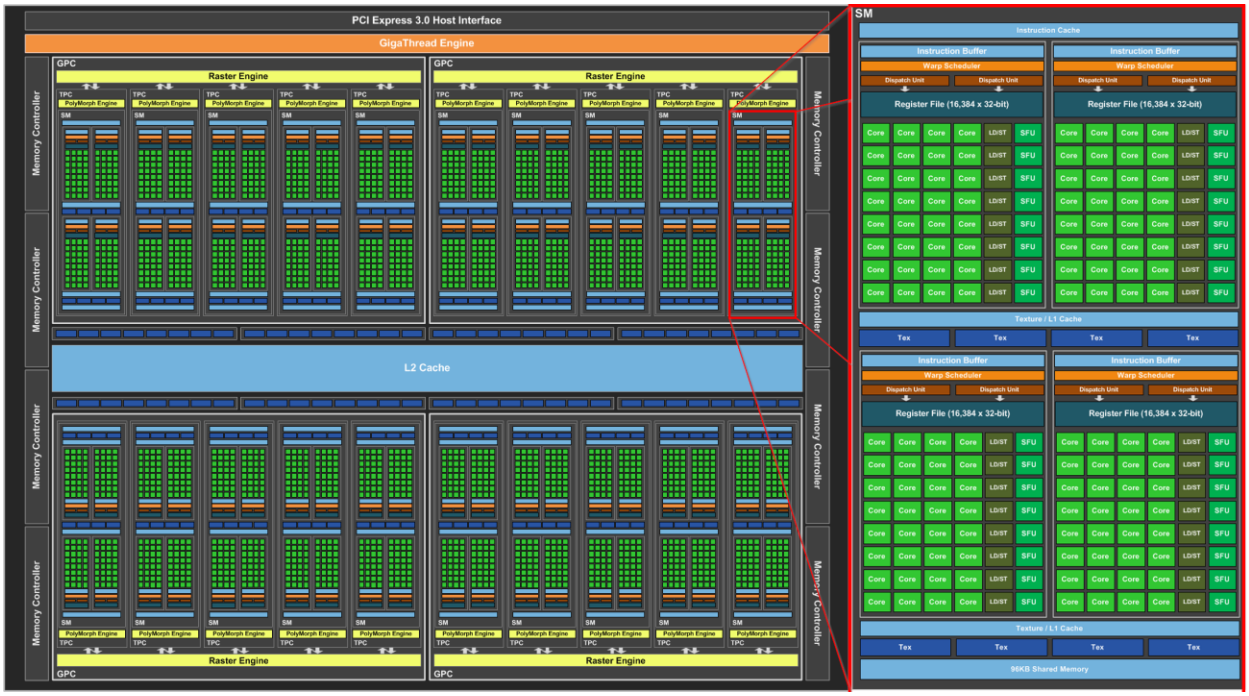


Specifically, the GP104 foregoes the HBM2 on-die ECC memory, the NVLink interconnects, the CUDA 6 Unified Memory architecture, and the high performance floating point capabilities of the GP100 / Tesla P100. It also seriously limits the FP16, or half precision, compute performance from GP100. In fact it is downright slow, signifying the differences between the HPC and consumer parts and their intended uses. While half precision performance is required for the kinds of things GP10 is built for (e.g. Deep Learning), it's not so much in demand on the desktop for mundane graphics demands.

Of these features removed for the desktop, the loss of Unified Memory is perhaps the most disappointing. Unified Memory on the Tesla GP100 enabled 49-bit virtual addressing, which is large enough to cover the 48-bit virtual address of modern CPUs plus the GPU's own memory, allowing one GPU to see all system virtual memory as well as GPU memory as one big memory space.

Internally, the GP104's structure looks very similar to Maxwell, but with more of just about everything. It still has Maxwell's Streaming Multiprocessor (SM) design which encapsulates four smaller units with 32 CUDA cores each (totaling 128 cores per SM), just as with Maxwell's GM204 which powered the GTX 980 and 980 Ti last year. Five SMs are housed inside of a Graphics Processing Cluster, or GPC, instead of the four SMs per cluster in Maxwell. There are four GPCs, so the GP104 has a total of 4 GPCs * 5 SMs * 128 CUDA cores/SM = 2560 CUDA cores total, up from 2048 in Maxwell's GP204.

The big story in Pascal along with the process shrink is the increase in raw speed. Base clock speed got a big boost in Pascal, up from 1126 MHz to 1607 MHz, with a boost clock of 1733MHz. This produces almost 9 TFLOPs of execution performance (up from about 5 TFLOPS in Maxwell) and a texture fill rate of 277.3 Gigatexels/s, up from 155.6 Gigatexels in Maxwell.



GP104 Pascal GPU Block Diagram

Details of the GP104 Pascal architecture can be found in the Nvidia GeForce GTX 1080 white paper here: http://international.download.nvidia.com/geforce-com/international/pdfs/GeForce_GTX_1080_Whitepaper_FINAL.pdf

Traditionally with the Maxwell, Kepler, and Fermi architectures, the initial Gx104 GPUs tended to act like a new architecture's training wheels, striking a balance between size and performance, and allowing Nvidia to get a suitably high yielding GPU at the start of a generation. These Gx104 GPUs would then be

followed up with larger GPUs later on as yields improve. With the GP104, Nvidia comes out of the gate with a fully mature design intended to run the most powerful graphics card models. It may be possible that a new GP2xx series is developed at 14nm that gives rise to a new series of graphics cards, but that's purely speculation at this point.

Pascal Iterations and GeForce Models

A particular graphics card's performance revolves around the layout and implementation of multiple SMs in the GPU. A full GP104 implementation as shown above is reserved for the high end GTX 1080 gaming card with 8GB of GDDR5X RAM. Because not everyone needs to have the best performing GPU, lower end GTX 10x0 cards will use close relatives to the GP104 but will have executions units disabled in order to differentiate performance and price points for different market segments. By the way, this is not just something Nvidia does. AMD makes their high end GPU then cuts it down to create lower end cards too.

Just below the GTX 1080 is the GTX 1070 with 8GB of GDDR5 RAM. Like the GeForce GTX 970 before it, it comes with the same GPU as its big brother, with internals disabled / detuned to bring it down to a more affordable market level. In this case the GP104 GPU has one entire GPC disabled, losing 5 SMs and 640 CUDA cores in the process for a total of 1920 cores. Nvidia also ratcheted down the GPU's base frequency to 1506 MHz and boost clock to 1683 MHz .

Further down the line we have the GTX 1060, GTX 1050, and GTX 1050 Ti. The GTX 1060 employs the GP106 GPU with 1,280 CUDA cores and has either 6GB or 3GB of GDDR5 RAM depending on model. The GTX 1050 Ti uses the GP107 CPU, with 768 CUDA cores housed in 2 GPCs of 3x128-core SMs each, and with 4GB of GDDR5 RAM. The GTX 1050 is a dialed down version of that, with one SM in the GP107 disabled for a total of 640 CUDA cores and only 2GB of GDDR5 RAM. Interestingly, both the GTX 1050 and GTX 1050 Ti use a smaller, 14nm FinFET process whereas the larger Pascal iterations use 16nm.

On the other end of the scale, we jump above the GTX 1080 with the GeForce GTX Titan. Powered by a GP102 GPU, it is architecturally similar to the GP104, just simply bigger. The four Graphics Processing Clusters in GP104 became six in GP102, each with 5 SMs, so we have a total of 30 SMs of 128 CUDA cores each. While that would normally provide 30 x 128 = 3,840 total CUDA cores, Nvidia disables two of the GPU's SMs to improve yields, bringing the boards' core count down to 3,584. With the additional cores, the Titan X also lowers the base clock to 1531 MHz, but overall TFLOPS score is still above 10 compared to 8.2 in the GP104. With 12GB of onboard memory, the Titan X is billed as the fastest desktop GPU ever, and at $1,200 it had dang well better be.

Workstation Class Pascal: The Quadro P Series

Nvidia recently announced some details on the Pascal versions of their Quadro professional graphics cards, the P5000 and P6000. As these details were released in October of 2016, we still do not have all of the facts (such as list price), but these are essentially "fully enabled" GP102 and GP104 cores. Basically the same guts as the GTX Titan X and the GTX 1080 without anything disabled.

The flagship Q6000 is a fully realized GP102 with 24GB of RAM, 3,840 CUDA cores and 12TFlops of computing power. The P5000 is thus similar to the GTX 1080, with the same GP104 GPU. However, the GTX 1080 was already "fully enabled" so the advantage of the P5000 over the GTX 1080 is mysterious at best, especially when you consider the multi-thousand dollar difference between what is effectively the exact same thing.

Nvidia GPU Roundup Comparison Chart

Below is a comparison chart which enumerates all of Nvidia's latest Pascal based gaming and professional graphics cards:

Nvidia Pascal GPU Comparison Chart						
	GTX Titan X	**GTX 1080**	**GTX 1070**	**GTX 1060**	**Quadro P6000**	**Quadro P5000**
GPU	GP102	GP104	GP104	GP106	GP102GL	GP104GL
CUDA Cores	3584	2560	1920	1280	3840	2560
SMs Enabled / Total	28/30	20/20	15/20	10/10	30/30	20/20
Core Clock (MHz)	1417	1607	1506	1506	1417	1607
Pixel Rate (GPixels/s)	136	102.8	96.4	72.3	136	102.8
Texture Rate (GT/s)	317.4	257.1	180.7	120.5	340	257.1
Single Precision (GFlops)	10157	8228	5783	3855	10883	8228
Memory Config	12GB GDDR5X	8GB GDDR5X	8GB GDDR5	6GB GDDR5	24GB GDDR5X	12GB GDDR5X
Memory Bus Width	384-bit	256-bit	256-bit	192-bit	384-bit	256-bit
Memory Bandwidth (GB/s)	480	320	256	192	480	320
TDP	250W	180W	150W	120W	250W	180W
DirectX/OpenGL/SM	12.0/4.5/5.1	12.0/4.5/5.1	12.0/4.5/5.1	12.0/4.5/5.1	12.0/4.5/5.1	12.0/4.5/5.1
Power Connectors	1x 6-pin + 1x 8-pin	1x 6-pin + 1x 8-pin	1x 6-pin + 1x 8-pin	1x 6-pin + 1x 8-pin	1x 6-pin + 1x 8-pin	1x 6-pin + 1x 8-pin
Outputs	1x DVI-D	1x DVI-I	1x DVI-I	1x DVI-I	1x DVI-D	1x DVI-D
	1x HDMI	1x HDMI	1x HDMI	1x HDMI	0x HDMI	0x HDMI
	3x DisplayPort	3x DisplayPort	3x DisplayPort	3x DisplayPort	4x DisplayPort	4x DisplayPort
Newegg Price (Nov. 2016)	$1,124.09	$649.00	$429.00	$249.00	?	?
Performance Ratio	100%	85%	72%	54%	100% +/-	85% +/-

Notes on Nvidia's Pascal GPUs

1. To save space the GTX 1050 and GTX 1050 Ti are not represented in the chart above. The GTX 1050 retails for around $140 and the GTX 1050 Ti is about $160.
2. The GTX 1050, GTX 1050 Ti, and GTX 1060 can be considered entry level cards for AEC applications, but are rather underwhelming compared to much more capable models with many more CUDA cores. On the other hand, I consider these excellent cards – particularly the GTX 1060 with 6GB – for Revit grunts and folks who otherwise do not use a lot of other exotic 3D applications.
3. The GTX 1070 can be considered the best overall bang for the buck. It provides about 85% of the performance of the GTX 1080 but is only 66% of the price.

4. Overall the GTX 1080 has the highest specifications for the mainstream market, and is considered the realistic performance king. Its specs are slightly better than the P5000 (which may be thousands more) and it is mid-priced at $650. It is built on the full-blown GP104 GPU with all 2560 CUDA cores enabled. It is also clocked at 1607 MHz, the highest of any card in this review and comes with 8GB of RAM.
5. The Titan X proves you don't need a $6,000 card to be the biggest kid on the block. Its only problem is the GTX 1080 is so close in performance and available for about half of the price.
6. The P5000 is in a terrible spot. Essentially the same card as the GTX 1080, it really doesn't have any reason to exist, if the GTX 1080 provides similar capability in all high-end applications. Being a Quadro, it has certified Quadro drivers which may mean something depending on your application.

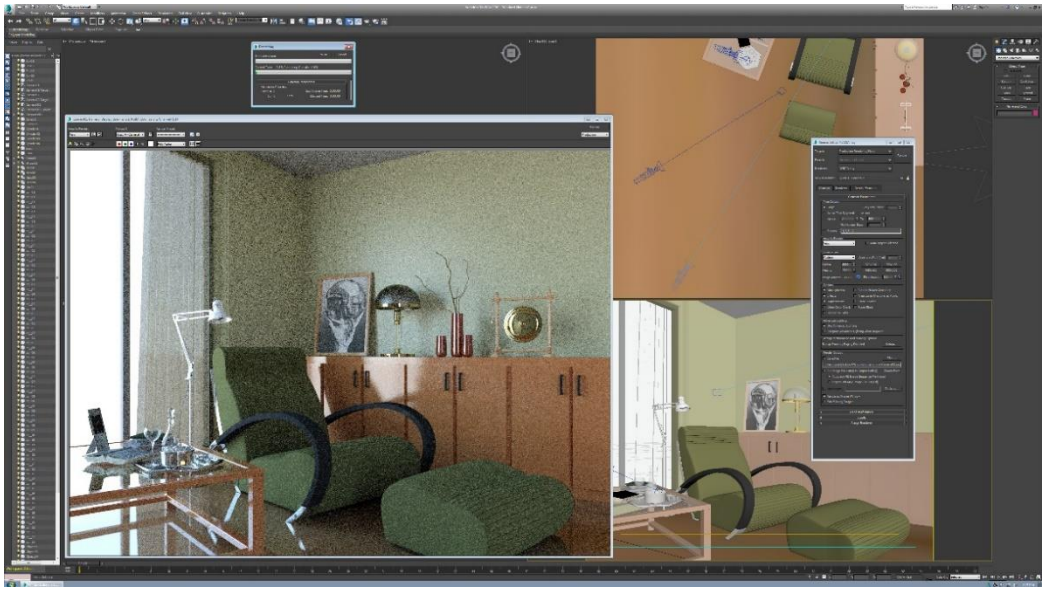
Gaming vs. Workstation Graphics Card Benchmarks, Cost Comparisons and Other Factors

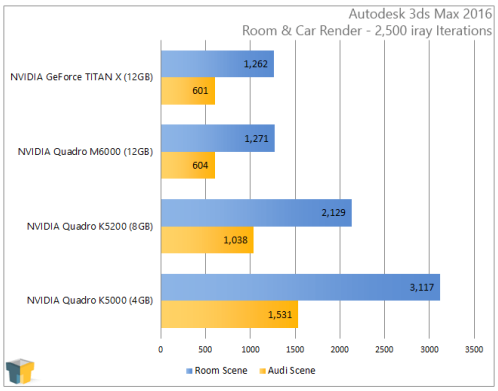
When comparing workstation class cards to their gaming card cousins, it's important to look at performance vs. price and make a judgement as to whether spending more on a workstation class card makes solid financial sense. Luckily, this is very easy to do and reveals some startling fact.

Because we don't have final prices on the Pascal Quadro cards, we will compare Maxwell-era technology, but the relative price / performance is similar. Recall the $50 difference between the middle-of-the-road GTX 970 gaming card and the more expensive but less capable K2200. Now let's look at the super high end, explicitly for CUDA / Iray / V-Ray RT rendering: Last year's top of the line Quadro M6000 retails for an eye popping $5,000. It runs the full implementation of the Maxwell GM200 CPU, with all 3,072 CUDA cores and all 24 SMM units operational, at a core clock of 988 MHz, and has 12GB on board. Being that CUDA performance scales linearly, we can look at the $/CUDA core as a metric. At $5K the M6000 has a metric of $1.63/CUDA core.

Then let's look at a flagship "gaming" version of this, the GeForce GTX Titan X also based on the 2nd generation Maxwell architecture. Also using the full implementation of the GM200 GPU, it has 24 SMMs and 3,072 CUDA cores, clocks at a slightly faster 1000 MHz and also has 12GB of video RAM. With a price of $1,000, the GTX Titan X cost 1/5 that of the M6000, and has a CUDA metric of only $0.33 / CUDA core.

Comparing a series of benchmark scenes in 3ds Max 2016, the GTX Titan X ***beats*** the M6000 by a small amount, thanks to the slightly higher clock speed. And both beat the legacy Kepler cards by a mile.





https://techgage.com/print/maxwell-hits-the-workstation-Nvidia-quadro-m6000-graphics-card-review/

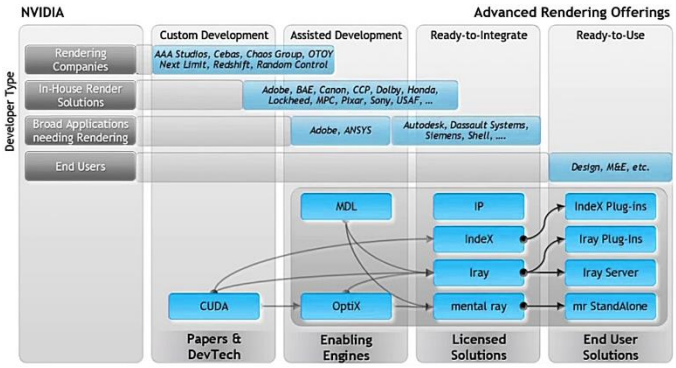
In other words, you could conceivably install a Maxwell-based GTX Titan X, get the same or better results as an M6000, run it until it blows up, get another one, run it into the ground until that one explodes, and repeat the process 3 more times before you've spent the equivalent of one M6000 card. Or, even better, you could install 4 GTX Titan Xs in a single system and get 4x the performance and STILL save $1000.

GPU Rendering: The Iray Factor and CUDA Technology

The choice of video card narrows and the picture focuses much more clearly if you want to render using a GPU-accelerated rendering engine such as the Iray or V-Ray RT rendering engines in 3ds Max Design.

As we discussed in the first section on Industry Pressures and Trends on page 10, there are several technologies out there (or on their way to the market) which leverage the computing power of the GPU for certain operations. Nvidia's CUDA platform and API helps turn your GPU into a very powerful CPU for parallel operations, such as rendering, performing audio analysis, geospatial applications, deep learning (i.e., self-driving cars), and a vast array of others. Nvidia also develops an enabling engine called Optix which helps 3rd parties to write renderers much easier that trying to do so from scratch. Additionally, Nvidia created a new Material Definition Library, or MDL, which works to create even more realistic materials easily in a standard way.

Both Iray from Nvidia and V-Ray RT from Chaos Group implement CUDA and Optix technologies. They bring GPU enabled rendering plugins to a wide array of Autodesk applications, such as Revit, 3ds Max, Maya, and Softimage, as well as others such as SketchUp, Rhinoceros, MODO, NUKE, Cinema 4D, Blender, and Katana. Both are available as standalone renderers as well.



Iray, for its part, is developed by Nvidia's Advanced Research Computing (ARC) group, as is the old standby rendering engine mental ray. Traditionally, 3D visualization artists made heavy use of mental ray, which was acquired by Nvidia when it bought mental images in 2007. Autodesk licenses both of these engines, takes the base code from Nvidia and builds application-centric user interfaces around it. This is why mental ray rendering in Revit was much simpler than in 3ds Max, because much less of the renderer controls were exposed to the end user in an attempt to make it more of a pushbutton operation.

With 2017, Autodesk removed the mental ray engine from Revit in favor of its Autodesk Raytracer Renderer (ART), which is much simpler, usually faster, and arguably produces better images. But neither ART or mental ray use the GPU at all – they are CPU only and are extremely slow.

Conversely, Iray was designed from the ground up to do two major things. First, make it pushbutton simple to create stunning photorealistic imagery. Physically accurate lighting and materials are designed to work together to produce stunning results – so good that it can be used for all sorts of real-world analysis. Instead of fiddling with literally hundreds of little knobs and controls all throughout the interface to fake realism, the idea is to model up the scene, install the lighting, set up materials, and push the Render button. That's it. Come back a little later and you have yourself a very, *very* good photorealistic image that arguably beats what you can get out of mental ray and is far easier to boot.

One of these is a photograph, the other is an Iray rendering. I can't remember which one is which.

Second, Iray was designed to take advantage of the new CUDA GPU hardware platform and programming model found in Nvidia graphics cards (and only Nvidia graphics cards) to perform the rendering on the card itself, not on the CPU. A single GPU is typically 4-12x faster than a quad-core CPU for these kinds of parallel tasks, and will complete these kinds of images in record time.

Not only does it use the GPU to accelerate rendering, but it scales very linearly with the number of GPUs in your system. This is a primary reason why higher-order HEDT or Xeon based systems are more compelling, because their additional PCIe 3.0 lanes allow more than one GPU to be installed. This makes it easier than ever to create a rendering workstation powerhouse without spending thousands more on exotic high-core-count CPUs. We cover PCIe 3.0 lanes later in this section.

Both Iray and V-Ray RT are becoming available as a plugin rendering engine for Revit with advanced capabilities. Nvidia currently has the beta of BIMIQ, an Iray rendering engine plugin for Revit, available at www.bimiq.com. V-Ray has a beta of its plugin for Revit which includes V-Ray RT GPU rendering capability.

Rendered in Revit using the BIMIQ Render Iray engine plug-in (left) and V-Ray for Revit interior scene (right)

Notes on Pascal and Iray

At the time of this writing (late October 2016), Iray is not compatible with the Pascal architecture and none of the graphics cards mentioned above will work under Iray in 3ds Max.[28] Pascal support is a priority with Nvidia, obviously, but because the P-series Quadro was just announced, Nvidia has to play a little catchup to get Iray fully functional on it. Look to see a service pack or hotfix for 3ds Max sometime soon which brings full Pascal compatibility to Iray.

What about AMD?

AMD produces great graphics cards, **and most middle to high-end Radeon HD cards will work perfectly fine in Autodesk's AEC applications**.

However, AMD does not implement the CUDA architecture, so their GPGPU capabilities are limited to using OpenCL, an open-source framework for writing applications that execute on GPUs as well as CPUs. Because of this and for the sake of brevity, I have not included AMD cards for analysis.

The bad news is that not many folks develop GPU renderers for OpenCL. With V-Ray RT and Iray being such popular GPU renderers, and the Optix plumbing available for others to easily create GPU accelerated renderers, there is little impetus for someone to develop the kind of widespread OpenCL libraries you need to gain traction.

However, OpenCL is supported by high end renderers like Indigo (http://www.indigorenderer.com) that can plug into 3ds Max, Maya, Revit, SketchUp, Blender, and others. OpenCL acceleration performance is a hallmark of the AMD Radeon HD series of cards – and they absolutely blow Nvidia cards out of the water in OpenCL specific benchmarks; see http://bit.ly/1BgJBdi

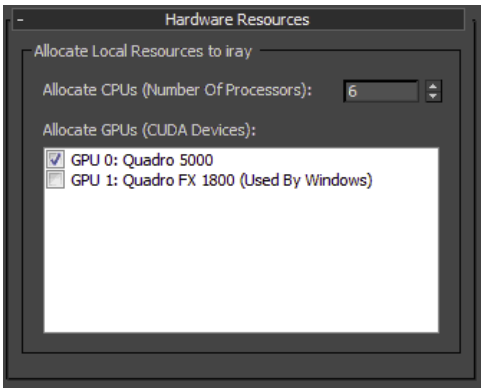
Multiple GPU Considerations

The thought of putting more than one GPU in your workstation may at first seem a little nuts, but the truth is that it is becoming the norm for people who perform GPU rendering on a regular basis.

You may have heard of existing technologies for running multiple GPUs together to aggregate performance, such as SLI (Nvidia) and Crossfire (AMD). These are technologies meant for gamers who can tie two cards together and combine their pixel pushing power for maintaining high frame rates. **Neither SLI nor Crossfire works for Iray or any other Autodesk product.** In fact, enabling SLI actually ***decreases*** Iray performance quite a bit.

When you render on Iray on a single GPU system, the graphics card will be running full tilt until it is finished. That means any movement by the mouse or trying to switch applications in Windows will temporarily stall the rendering process and system responsiveness will grind to a halt. Most people doing GPU rendering have learned early on to install at least one other card in the system just to handle the Windows OS operations. This doesn't have to be a very high end card - in fact, any decent card will do. In 3ds Max Design's Iray renderer rollout, you can assign the specific card you want to dedicate to handling Windows and which one you want to handle the rendering task. You can also instruct it to limit the number of available CPU cores as well to maintain system responsiveness.

[28] http://blog.irayrender.com/post/150442269131/i-was-wondering-if-iray-is-being-depreciated-iray



The other consideration is how many physical cards you can fit into a single workstation. This is limited by the number and type of PCIe slots you have available. All systems have a limited number of PCIe slots, some may run faster than others, and your computer's case may be too small or not ventilated well enough to support more than two full length cards at one time.

Or, you may have a laptop that obviously cannot be expanded. For these cases, you can invest in external PCIe enclosures which can house 1 to 4 or more PCIe cards. These additional cards aren't connected to any monitors - they are simply graphics coprocessors which are churning on your renderings. You can interface this with your computer via an adapter card or a Thunderbolt connection.

The main reason to use multiple GPUs is that Iray rendering performance scales almost directly with the total number of CUDA cores available. Put simply, for a single card the number of CUDA cores x the base clock frequency = relative performance. Additionally, Iray can leverage and automatically use all of the CUDA cores across separate video cards. Add more CUDA-compatible GPUs to a system and watch your render times drop, and fast.

GPU scalability is an important factor to analyze when specifying graphics cards. Is it more cost effective to use one very large and expensive GPU, or have several lower-end ones that all work together?

Here is a chart showing how multiple GPUs in a single system scale up pretty linearly. 2 GPUs is roughly 2x faster; 4 GPUs is 4x faster, etc., so the number of CUDA cores compared to the total price to get to that number is an important metric. Given the relatively affordable prices for GTX 1080 and GTX 1070 cards and the way the price curve jumps very high with Titan X, it perhaps makes sense to scale performance by putting more of them in your case instead of a faster one.

GPU	CUDA Cores	Total Cost
GTX 1080	2560	$650.00
GTX 1070	1920	$429.00
GTX 1060	1280	$250.00
(2) GTX 1080	5120	$1,300.00
(3) GTX 1070	5760	$1,287.00
(4) GTX 1060	5120	$1,000.00

Here you can see that, when calculating value in terms of $ / CUDA performance, an affordable card like the GTX 1070 really can't be beat in a multi-GPU setup. It provides more CUDA cores for less money than the more expensive dual-GTX 1080 setup and therefore better GPU rendering performance. The only better value is (5) GTX 1060s, which nets you $6,400 cores for $1,250. But no one really has 5 PCIe slots available, so the point is rather moot.

Inside PCI Express 3.0

Peripheral Component Interconnect Express, or PCI Express for short, or **PCIe** for shorter, is a high-speed serial expansion bus standard that replaced the older PCI, PCI-X, and AGP standards. Unlike the old standard slots dedicated for graphics, a PCIe standard slot can host any kind of compatible expansion card, whether it is for graphics, sound, storage, networking, or whatever. In fact there is a large push today to move Solid State Drives from SATA to the PCIe bus directly, which will be discussed in the next section.

PCIe is based on a point-to-point serial bus topology that uses serial links (point to point communication channels) connecting every device to the host controller, which is either on the CPU itself or on the motherboard's chipset. A single link between two devices is comprised of 1 to 32 lanes, which are the physical signaling traces that run to each PCIe slot. (Each lane is actually composed of 4 wires or signal traces on the motherboard). PCIe defines slots and connectors supporting multi-lane links from one to 32 lanes in powers of 2 (1, 2, 4, 8, 16, or 32) increments. Links are expressed with an 'xN' prefix, so x16 ('By 16') represents a 16-lane card or slot which is the largest commonly used size.

Given this, slots come in different sizes given in standard x1, x4, x8, x16 notation, which represents the largest physical card that will fit. The lane size is automatically negotiated during device initialization, so smaller cards can fit in larger form factor slots; e.g., you can put an x8-size card in an x16-size slot without issue and it just works.

However, size alone does not necessarily refer to the slot's bandwidth; that is determined by the lanes going to the slot regardless of size. When there is a difference this is usually specified as "xsize @x speed," for example, "x16 @ x8" means an x16 size physical slot that is configured with only 8 lanes. Pay attention to your motherboard manual to understand which slot can do what. On some motherboard, the PCIe slots are colored to indicate their capacity, but that is solely up to the manufacturer / model.

When considering a system that you know you are going to install multiple GPUs, review the motherboard specifications. You will often see the PCIe expansion slots listed as follows: →

Expansion Slots

PCI Express 3.0 x16	2(x16 or dual x8)
PCI Express 2.0 x16	1(x4 mode)
PCI Express x1	4

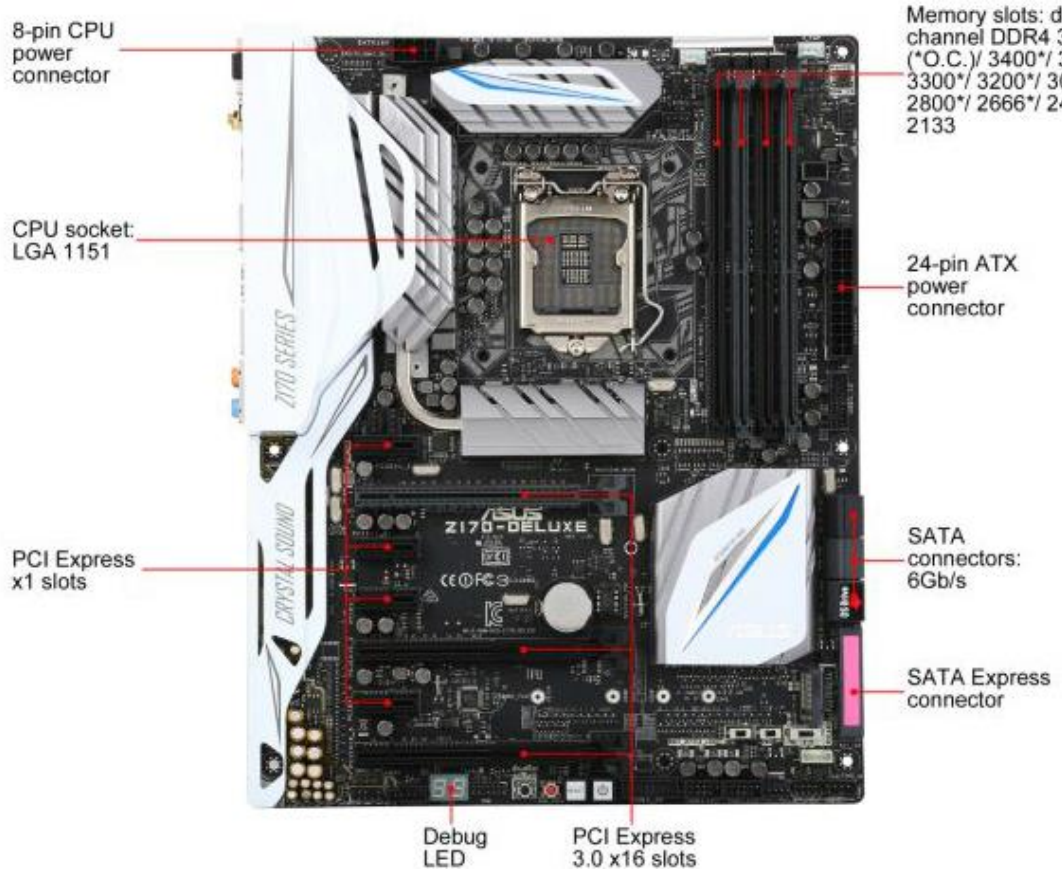
This refers to the number of physical PCI Express slots provided, which ones are PCI Express 2.0 and PCI Express 3.0 compliant, how large the slots are (i.e., how wide are the lanes serving it), and how the slots will use the available lanes in different configurations. Much of this is actually determined by the chipset used, which in turn is reliant on the CPU so there is not a lot of usable flexibility here per CPU platform.

The history of PCIe introduced several standards, each of which improves throughput and internal optimizations for signaling and data integrity over the previous version. PCIe 1.0 had an effective bandwidth of 250 MB/s maximum data rate per lane, resulting in a 4 GB/s aggregate rate per x16 connector. PCIe 2.0 bumped this up to a 500 MBs / lane rate (8 GB/s for an x16 connector) and is still commonly used today. PCIe 3.0 doubled this to 984.6 MB/s per lane (almost 16 GB/s per x16 connector) and is the latest standard. The PCIe 3.0 standard was first fully supported by the Ivy Bridge CPU introduced in 2012 and is used by all graphics cards today. PCIe 4.0 is expected to double this again to 1969 MB/s per lane or 31.5 GB/s per x16 slot.

As noted in the section of processors, CPUs have their own embedded PCI controller which provides anywhere from 16 to 40 PCIe 3.0 lanes depending on model. These lanes service the x16 expansion slots used for graphics cards, allowing for up to dual GPU configurations for 16-lane GPUs, and support up to 5 graphics cards on a 40-lane GPU.

The latest Z170 PCH, with its DMI 3.0 connection to the CPU, provides 20 PCIe 3.0 lanes for greater peripheral connectivity such as networking, audio, and storage. In fact, storage is one reason why Intel upgraded DMI 2.0 to DMI 3.0, effectively doubling the bandwidth between the CPU and PCH. As discussed in the next section, SATA as an interface to high end SSDs is coming to an end, and PCIe based SSDs are going to roam the Earth next year. The expected massive increase in SSD bandwidth requires a faster connection to the CPU.

While the total number of lanes is a physical constant, the configuration of PCIe slots is always a compromise between throughput and number of devices at that speed. The motherboard shown below is an ASUS Z170 Deluxe LGA 1151, typical of current Skylake systems. The allotment of PCI lanes across the 7 PCIe slots is a variable depending on what is plugged in where.



7 PCIe slots divide up the allotment of PCIe lanes in this manner:

- There are two PCI Express 3.0 x16 slots that will run a single GPU at x16 or dual GPUs at x8/x8. These PCIe slots use the 16 lanes from the CPU.
- 1 PCI Express 3.0 x16 slot which runs at maximum x4 mode, compatible with PCIe x1, x2 and x4 devices. However, this PCIe x16 slot shares bandwidth with SATA connectors. The PCIe x16 is by default set at x2 mode.
- 4 PCI Express 3.0 x1 slots, typically for audio cards, modems, or other low-speed adapters.

You can see above that you need to pay particular attention to the configuration of the PCIe slots on the Skylake platform and make sure you plug in the right things to the right slots, or you end up with something that needs more bandwidth than the slot can provide. For example, Nvidia cards require at least a PCIe 3.0 x8 slot to even work. Put it into a x16 slot running x4 and nothing happens. Furthermore, often if you plug something into one slot, another slot will go from x16 to x8 speeds because the lanes are all shared. The motherboard manual will go into excruciating detail on what you can plug into what slot.

PCIe 2.0 vs PCIe 3.0

It is interesting to note that the throughput provided by PCIe 2.0 was underutilized by even the most powerful graphics cards, so the doubling of bandwidth in PCIe 3.0 by itself provided little to no improvement performance when comparing the same card on both slots. Rather, PCIe 3.0 provides plenty of headroom for multiple GPUs to be installed in the same system and each use more available lanes. Because PCIe 3.0's throughput is roughly double that of PCIe 2.0, the throughput for a PCIe 3.0 card at x8 will equal that of a PCIe 2.0 card running at x16, which is fine. Even running at x8 mode, a powerful graphics card is not bandwidth limited by the PCIe bus, so there is no problem in running two PCIe 3.0 x16 cards at x8. In fact, running a PCIe 3.0 x16 card at x4 speed (at 25% of the available bandwidth) reduces overall performance by only about 14%.

Maximizing GPU Rendering on Advanced Platforms

Because of the 16 PCIe lane limitations of Skylake and Xeon E3 CPUs, high-end Viz Wizards who need more than two GPUs necessarily need to move to a more advanced CPU / motherboard platform. With 28-40 available PCIe 3.0 lanes provided by the Broadwell E and Xeon E5 platforms, users of GPU renderers have the ability to use 3-way, 4-way, or even 5-way PCIe 3.0 cards, lowering render times in Iray and increasing viewport interactivity by leaps and bounds.

For example, the ASUS x99-Deluxe motherboard for the Broadwell E platform (which can support CPUs with either 28 or 40 PCIe 3.0 lanes) has five x16 slots and one x4 sized slot. The expansion capabilities are:

PCI Express 3.0 x16
40-Lane CPU: 5 x PCIe 3.0/2.0 x16 (x16, x16/x16, x16/x16/x8, x8/x8/x16/x8, x8/x8/x8/x8/x8 mode)
28-Lane CPU: 3 x PCIe 3.0/2.0 x16 (x16, x16/x8, x8/x8/x8)
Note: The 5th PCIe x16 slot shares bandwidth with M.2 x4. Triple PCIe 3.0/2.0 configuration is default set at x8/x8/x8. Adjust PCIEX16_5 Slot Bandwidth in BIOS.

PCI Express 2.0 x16
28-Lane CPU: 2 x PCIe 2.0 x16 (x1 mode)

PCI Express x4
40-Lane CPU: 1 x PCIe 2.0 x4 (max at x4 mode)
28-Lane CPU: 1 x PCIe 2.0 x4 (max at x4 mode)
Note: The PCIex4_1, USB3_E12 and SATAEXPRESS_E1 connectors share the same bandwidth. The SATAEXPRESS_E1 will be disabled when there is a device installed on PCIEX4_1 slot. Set this option to X2 Mode or X4 Mode when the installed PCIe device is higher than X4 interface.

VII. Storage

Solid State Drives

Today, Solid State Drives (SSDs) are now considered mainstream components and are a necessity for any AEC application user, regardless of where they fall in the user profile range discussed in Section 1. Having no moving parts, all of the data is held in solid state non-volatile memory. There's no noise or heat, and they draw much less power than a mechanical hard disk.

The performance improvement of an SSD is truly dramatic over typical hard drives. **An SSD is arguably the single best upgrade you can make to any system of any speed**. Random access times are about 0.1ms, compared to 5-10ms for hard drives. Read times are not affected by file fragmentation or where the data is stored, compared to HDDs where data written to the inner cylinders are read more slowly than the outer cylinders. Driver support under Windows - especially Windows 8 and 10 - is excellent.

In particular, large applications such as in the AEC Collection take on a new life under SSDs. Every AEC application takes a while to load on mechanical drives, but are very quick to open from an SSD. The same goes for loading up large Revit models; the speed at which as SSD can deliver a 300MB+ file into system memory is pretty amazing.

Most SSDs are still using the old technologies that powered mechanical drives. Under the hood there are several newer, more appropriate technologies that serve Solid State Drives that merit discussion and consideration when specifying new workstations that rely on fast storage capabilities.

The Death of SATA

You are likely used to plugging in all of your disks plug into the SATA ports on the motherboard. This Serial ATA III interface uses the Advanced Host Controller Interface (AHCI) as the "software" interface which allows advanced features such as hot plug and native command queueing (NCQ) to optimize throughput.

The problem is that today's SSDs are now **too** fast for the traditional storage subsystem in most PCs. SATA and AHCI were originally intended to host mechanical hard disks, and are unprepared to take on the throughput of modern SSDs. The SATA III 6Gb/s interface limits us to 500+MB/s reads and sub-500MB/s writes in real-world performance numbers. New SSDs have sequential read/write speeds upwards of 2.8 GB/s and 1.9 GB/s respectively, so the SATA interface itself is limiting how fast our SSDs can perform.

Thus, over the past few years there has been a push to get SATA out of the way by utilizing the PCI Express bus. Instead of trying to double the bandwidth of SATA – which would take forever to catch up with the progress of SSD technology - the PCIe bus already provides the bandwidth required.

In 2014 there were attempts to fuse SATA and PCIe two together in a backward-compatible connector called SATA Express, but this was never really widely adopted. It uses the same physical connector as SATA drives but uses PCI Express lanes rather than the SATA bus to boost transfer speeds. It is not that much faster than SATA, as it communicates through 2 PCIe lanes, limiting the interface to 2GB/s. The industry collectively made up its mind that, if the move is clearly to go to PCIe based storage interfaces, they will figure out how it should function on its own instead of trying to shoehorn PCIe and SATA together. 2015 was the breakout year for PCIe based SSDs using the NVMe protocol and the M.2 physical connector.

Exit AHCI, Enter NVMe

The main problem with today's SSDs is AHCI, the software protocol designed to specify the operation of SATA host bus adapters to work with mechanical hard drives in mind. It is more optimized for high latency rotating media rather than low-latency non-volatile storage in SSDs. As a result AHCI cannot take full advantage of SSDs, and with PCIe the potential bandwidth increases dramatically. It's like driving a race car on a dirt road. We need a new software interface optimized for both SSDs and PCIe.

Non-Volatile Memory Express, or NVMe, has been designed from the ground up specifically to capitalize on the low latency and parallelism of PCI Express SSDs, and complementing the parallelism of contemporary CPUs, platforms and applications. At a high level, primary advantages of NVMe over AHCI relate to NVMe's ability to exploit parallelism in host hardware and software, based on its design advantages that include data transfers with fewer stages, greater depth of command queues, and more efficient interrupt processing.

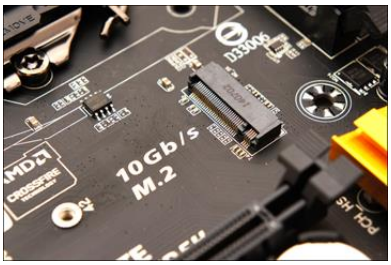
	NVMe	AHCI
Latency	2.8 µs	6.0 µs
Maximum Queue Depth	Up to 64K queues with 64K commands each	Up to 1 queue with 32 commands each
Multicore Support	Yes	Limited
4KB Efficiency	One 64B fetch	Two serialized host DRAM fetches required

Source: Intel

Say Hello to your New Little Friend – M.2

A faster interface is not the only thing that is changing on the SSD horizon. Meet your new hard drive:

The Samsung 960 Pro pictured above uses the new **M.2** interface, and at 80mm long it's a little over half the length of a DDR4 DIMM module. This new PCI Express Mini Card form factor plugs directly onto your motherboard and right into the PCIe 3.0 bus, using up to four lanes. Alternatively, inexpensive PCIe adapter cards are available if you have a free x4 PCIe slot.



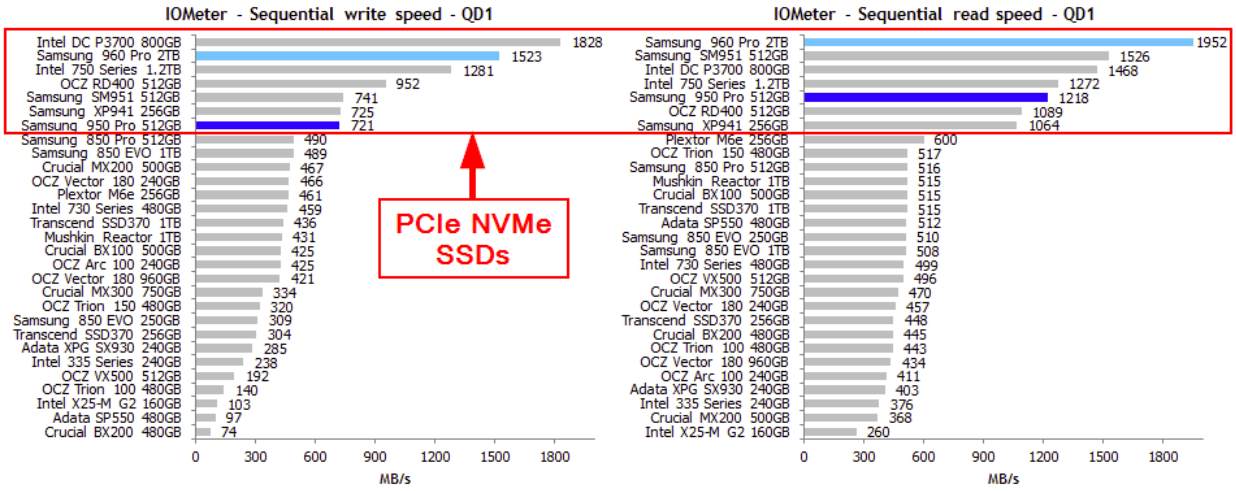
The new M.2 interface is really the successor to mSATA (mini-SATA), but it is far more versatile. M.2 is a connector form factor, and M.2 connectors can plug into different buses on the motherboard – PCIe, USB 2.0 and 3.0 buses, SATA III, DisplayPort, and others. There is more than one kind of M.2 connector, more than one type of interface that can be used with M.2, and more than one kind of M.2 card. To keep things straight, M.2 connectors can have different keying notches which denote different purposes and capabilities.

M.2 supports a wide variety of hardware, including Wi-Fi and Bluetooth combo cards, sat-nav cards, audio cards, and of course SSDs. M.2 is also great because it does not require separate power and data cables - in fact there are no cables, it's just a motherboard connector, usually sandwiched between PCIe slots. M.2 supports NVMe as the logical device interface, while also maintaining backward compatibility with AHCI (at the cost of not delivering optimal performance).

M.2 cards can range from full length 80mm boards used for SSDs down to 42mm cards usually designed for wireless chips. The full M.2 experience, connecting to 4 PCIe lanes, is available on modern Skylake Z170 motherboards, as well as the Broadwell E X99 platform. Older CPUs and chipsets, e.g. Z97 for Haswell reserved only two PCIe lanes for M.2, limiting bandwidth to 2GB/s.

Performance Improvements with PCIe NVMe

As you can see in the benchmarks below[29], PCIe/NVMe drives trounce their backwater SATA/AHCI cousins by a huge amount. Even real-world benchmarks, which usually decrease the kinds of dramatic differences expressed in synthetic scores, show astonishing performance gains with PCIe/NVMe.



In the above synthetic benchmarks you see all of the PCIe NVMe drives well ahead of typical 2.5" SATA SSDs. The top performing Intel drives are enterprise-class drives on a traditional x4 PCIe card; the rest of the top performers are M.2. While enterprise-class drives are usually the fastest SSDs you can buy, they are also very expensive ($700 - $1,000) and require a PCI-e 3.0 x4 slot.

Sizing up SSD Sweet Spots - 512GB is the New Black

One advancement in 2016 is the standardization in 512GB as the minimum storage capacity that most will opt for. 512GB drive prices have dropped to the point that 256GB drives were a year ago, and 1TB drives aren't as stratospherically expensive as they once were.

256GB was *just* enough for your OC and applications, but not much else, so most people installed a small 256GB drive along with a larger mechanical drive for mass storage, which is inefficient. 256GB drives also perform worse than higher capacities, making them even less desirable.

Currently there is a price premium on PCIe NVMe drives, making them more expensive per GB than typical 2.5" drives of the same size, but that should go away as they become the new standard and SATA SSDs start to disappear.

When SSDs first came out en masse in 2012-2013, reliability was of great concern, because we simply didn't know how much you could read and write off of them before they started to die. The NAND technology behind non-volatile RAM will inevitably wear down over time. In 2015 TechReport.com did an SSD endurance experiment[30] and found that consumer-grade SSDs can take a serious data pounding, with some passing the 2 *Petabyte* write mark, which is far more data than users will ever need to write during the useful lives of the drives.

Or their own lives, for that matter: Most users will never write more than a few Terabytes per year. 2 Petabytes is quite a bit and equates to 2,000,000 Gigabytes. This is 50GB writes per day for over 100 years.

[29] http://techreport.com/review/30813/samsung-960-pro-2tb-ssd-reviewed

[30] http://techreport.com/review/27909/the-ssd-endurance-experiment-theyre-all-dead

VIII. Mobile Computing

When purchasing a laptop or mobile workstation, you have different sets of parameters to think about before you make a decision. Those decisions have to be made on how you plan to use it on a daily basis more than anything else. Mobile computing becomes a compelling lifestyle addition and typically trumps raw performance. It's very easy to become seduced by the freedom of sitting in a coffee shop working on a Revit model rather than sitting in some fabric covered box all day.

Portability vs. Power

As mentioned in Section IV on Processors, the first consideration in shopping for a laptop is to understand how you will use it. Is it to be your one and only daily driver, or are you working between a laptop and a desktop on a regular basis? The amount of time you need to carry around your laptop will definitely carry more weight - excuse the pun - in your decisions about form factors, power requirements, and other factors we will discuss.

This question is largely between portability and power, because you really can't get too much of both in one unit. Ultraportable 13" machines are small and light and lovely to travel with, but have low-powered CPUs and memory constraints which work for Word or Chrome; not so much in Revit with large models.

This is changing, particularly in the system thickness. Today's latest processors and graphics chips require a lot less power and put out less heat, so new laptop models are getting much thinner. Form factor is specific to the particular machine, so as you compare models, you need to consider not only the internal components, but pay attention to overall size and weight specifications.

Generally speaking, laptops that you would want to run AEC applications on all day start at 15" screen sizes up to 17", and heft is definitely a factor. The more powerful the machine, the bulkier and heavier it is. 17" laptops are quite heavy to lug around, and forget about using one on an airplane.

Screen Choices

A major factor in deciding the base form factor is the screen size and resolution. Choose wisely, because this largely determines almost everything else about the system as it drives the overall usability of the system. Within each model lineup, you have only a few options for screen resolution, anti-glare options, and so on.

Because UI elements are defined in pixels, it's a balance to find a resolution that provides a decent size for icons and other UI elements without being too tiny. For example, the default in a low-cost 15.6" screen may be 1366x768, which is just too coarse for design applications. The next step up is 1080p, or 1920x1080 resolution which many consider perfect for that size form factor.

You may be tempted to go with higher resolutions such as QHD 3200x1800, but beware of scaling issues. Windows and application user interface elements, such as ribbon and toolbar icons, text, and so on are measured in pixels. As pixel density gets higher (resolution increases while screen size stays the same), you get smaller and smaller UI elements making it difficult to use.

3200x1800 offer a lot of screen real estate if left at its native resolution, but even those with 20/20 vision will have problems adjusting to tiny UI elements. Windows 10 works a lot better than Windows 8.1 and much better than Windows 7 at providing a good scaling factor of 150% to handle such high resolutions, which helps make font, icons, symbols and windows easier to read.

However, you may hit on a few stubborn programs that yield less desirable results. First, the screen in everyday Windows usage may appear more blurry, because you are scaling things up 1.5 pixels in each direction, and there is no such thing as half a pixel. Applications may also not respond well to scaling and the UI may reorient things in an unexpected manner.

A new “4K” standard is double that of 1080p, or 3840x2160 resolution. This standard is actually usable because you can scale up your Windows elements by 200%, or an even 2 pixels, so clarity should not be a problem. You may still experience a problem in how your applications can scale their UI elements to work with such high resolutions.

Previous releases of Autodesk’s applications were optimized for monitor displays under 2000 pixels wide. 3ds Max 2017 and the 2017.1 updates available for both AutoCAD and Revit provide high resolution screen support and allow clear scaling of 200% and beyond.

The problem is, of course, that not everyone is on 2017 version of their AEC applications, particularly Revit which has no backwards compatibility and version requirements within projects. If you expect to be running older versions of your Autodesk software – not an uncommon occurrence – you will experience artifacts, blurry screens, tiny UI elements, and other issues if you choose a high resolution screen. Perhaps there is a slim chance that 2016 products would get a 4K update, but so not place money on that bet.

All in all these higher resolutions may cause more problems than benefits, especially when you consider that there was nothing really broken about a 1920x1080 resolution on a 15.6” screen. For these reasons I would not recommend a 4K panel for Autodesk’s AEC application users for this year.

The other consideration is screen technology. Most mobile laptop manufacturers offer several different screens: IPS (In Plane Switching) screens and Twisted Nematic (TN) panels are the ones most offered. If possible, go for the IPS displays – they are brighter, have more vibrant colors, and have wider viewing angles. They do cost more, but you are going to live with the display for the life of the laptop. It’s best not scrimp in the really important stuff like what you stare every day.

With Windows 10 being so popular, the choice of touch screens is possibly compelling. For ultraportables or hybrid designs such as the Microsoft Surface and Surface Book, that is a must have feature. For working laptops, perhaps not so much. And if you are like me, anyone touching my screen with their greasy fingers is subject to the death penalty, so this may not be a worthwhile option. Antiglare versus glossy screens are another entirely subjective and personal preference. Here, a field trip to the local big box store to look at various differences in screens may be in order.

Processor Choices

As noted in the section on CPUs, there are several to choose from, but the choices you have depends on form factor, manufacturer, and model. From the new 7th generation Kaby Lake low power models, to the mainstream 6th generation Skylake i7-6xxx processors, up to mobile Xeon E3-15xx models, there is something there for everyone. Common sense will largely dictate the solution, given the often prohibitive upcharge to go from one model to another. Unless you are going for an ultraportable (and not considering a machine for heavy AEC app usage anyway), you want a true quad-core model somewhere in the 2.6 to 2.8 GHz range with 8MB of L3 cache.

System Memory

With any mobile platform you are limited on the amount of RAM you can install, typically to only 2 slots in mainstream models and 4 slots in mobile workstation-class machines. Sometimes 2 of those slots are under the keyboard, making upgrades more difficult.

As a desktop replacement for an AEC applications user, I typically recommend 32GB of RAM in 2x16GB modules as a minimum in a workstation class machine. That adequately covers 99% of common workloads and you won’t have to pull out memory you cannot use if you ever want to upgrade to 64GB. Remember that 64GB is really more appropriate for someone who is working on very large models, works across many applications simultaneously, or does a lot of rendering, and not many people do this high level of production work solely on laptops.

The same warnings for purchasing RAM from system vendors applies to laptops as well as desktops. Dell and HP will readily charge hundreds of dollars for even paltry upgrades. If your vendor won't match prices and you know how to work a screwdriver, buy the laptop with the bare minimum, pick up your RAM elsewhere and install it yourself. It literally takes minutes on most laptops. Just don't lose the tiny screws.

Graphics Choices

Your choice of graphics card is going to be severely limited within any particular line, so it may end up driving your form factor and model decision. What goes for desktop cards goes for laptops; stay away from the integrated Intel HD graphics unless you rarely touch AEC applications such as Revit, 3ds Max, and Navisworks. That includes gaming as well.

Workstation class laptops are going to come with workstation class graphics, and there are not many viable options for most vendors. In any particular model, I typically recommend upgrading to the fastest option available within reason, because you never know what you may work on next year that really takes advantage of it. Do a quick ROI: a $300 graphics board upcharge works out to a little over $8 a month over the 3-year life of the laptop.

However, be cognizant of the differences between GPUs. On the Nvidia front, the Quadro GPUs available for a workstation-class 15.6" mobile platform are the M1000M and M2000M, based on the 1st generation Maxwell GM107 GPU. On the 17.3" platform you have the higher-order GPUs in the M3000M, M4000M, M5000M, and M5500,[31] all based on the 2nd generation Maxwell GM204 GPU. Note that the M5500 is based on the GTX 980 (Notebook) GPU which is actually a desktop GeForce GTX 980 shoehorned into a mobile form factor, and is usually found on extremely bulky laptops. It even supports overclocking.

Nvidia's Quadro offerings on mobile workstations follow the same rules as in their desktop cards: ultimately, the GPUs are very similar to those in GeForce cards. The major properties are shown here.

Nvidia's Mobile Quadro Comparison Chart						
	17.3" Platform				15.6" Platform	
	M5500	M5000M	M4000M	M3000M	M2000M	M1000M
GPU	GM204	GM204	GM204	GM204	GM107	GM107
Fabrication	28 nm				28 nm	
CUDA Cores	2048	1536	1280	1024	640	512
Core Clock (MHz)	861	975	975	540	1029	993
Pixel Rate (GPixels/s)	55.1	62.4	62.4	17.28	32.9	15.89
Texture Rate (GT/s)	110.2	93.6	78.0	34.6	41.2	31.8
Single Precision (GFlops)	3527	2995	2496	1106	1317	1017
Memory Size	8GB	8GB	4GB	4GB	4GB	2GB/4GB
Memory Bus Width	256-bit	256-bit	256-bit	256-bit	128-bit	128-bit
Memory Bandwidth (GB/s)	211	160	160	160	80	80
TDP	150W	100W	100W	75W	55W	40W
Mobile GTX Equivalent	GTX 980 (Notebook)	GTX 980M	GTX 970M	GTX 965M	GTX 960M	N/A

[31] http://www.nvidia.com/object/quadro-for-mobile-workstations.html

Note from the table above the performance overlap between the M2000M on the 15.6" platform and the comparatively horrendous performance of the M3000M on the 17.3" platform. Here the additional cores in the M3000M are no help to the very slow core clock speed which tanks its performance numbers.

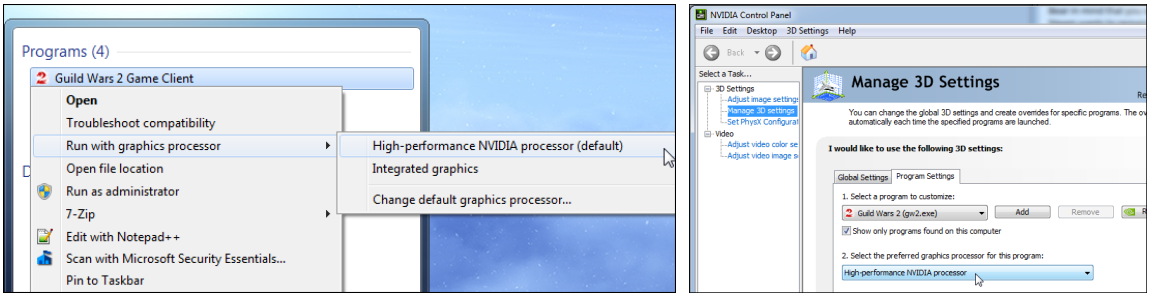
If you need dedicated Iray rendering on a mobile workstation, opt for the 17.3" platform and outfit it with a Quadro M5500. Just be prepared to pay an enormous amount of money for a very heavy laptop.

We do not yet have a mobile Pascal Quadro offering, but being that Pascal is a 14nm design and a much more efficient chip, I would expect to see it in come out in early 2017. GTX 10-series Pascal offerings are available in some consumer-based laptops, but they are largely boutique builders catering to gamers at the time of this writing. Companies such as CyberPowerPC, Falcon Northwest, and Origin cater to the enthusiast market and put out strikingly similar gaming laptop designs outfitted with Pascal GPUs. The problem is that they are extremely bulky in comparison to more commercial offerings from Dell, HP, and others. For these vendors you will see GTX 9xx and 8xx series GPUs which should function for most purposes, even in AEC applications, just fine.

Optimus

Nvidia's workstation mobile platform includes something called Optimus technology. Optimus is a GPU switching technology designed to "seamlessly switch between two graphics adapters within a computer system in order to provide either maximum performance or minimum power draw from the system's graphics rendering hardware."[32] This switching allows the system to power up and use the discrete GPU in applications that can use it, then revert to the integrated GPU (IGP) when you do not need it.

This happens in the driver, which tries to determine if the application would benefit from using the discrete GPU. Even in this case, the integrated GPU (IGP) still outputs the final image. When less demanding applications are used, the internal IGP takes over. You can override this in the Nvidia Control Panel driver settings, selecting which application uses which GPU. The Nvidia driver also provides the option to select the GPU on demand, in the right-click menu when launching the application.



However, this interplay is tricky and may not work all of the time, particularly with Autodesk applications. Under Windows 7 and 8 I've seen Optimus cause serious workability issues in Revit which make it difficult if not impossible to use. If you are having graphics related issues in an Nvidia-based workstation, first ensure both the Nvidia and Intel video drivers are up to date.

Failing that, look for the settings for Optimus technology in the BIOS and turn it off, using the high performing Nvidia GPU 100% of the time. Note that that this option may not be found in the BIOS of non-workstation class laptops. If you travel a lot and work off the battery and/or only need the discrete GPU sometimes, then keep it on and see if it works for you.

[32] https://en.wikipedia.org/wiki/Nvidia_Optimus

Storage

In a typical mobile workstation, storage is going to be limited to a few choices. Some systems may offer storage as a single 2.5" drive or, more likely, in an M.2 slot, or in combination with a mechanical drive that is housed in the optical drive bay. You may also have the option for a 1.8" "Mini-Card" that can typically be a SSD, so your choices may allow for one SSD only, two SSDs, or an SSD and a 2.5" mechanical disk. As with desktop systems, ensure you get an SSD as your boot drive in all cases.

Regardless of use, I recommend you opt for a single 512GB SSD instead of a smaller 256GB SSD and an additional drive. Remember that a second mechanical drive will consume power and heat and lower battery life, as well as add weight.

If you think you need to store more than 500GB worth of OS, apps, and data on it, go for either a 1TB SSD or a secondary mechanical disk. For mechanical drives, always opt for a 7200 RPM drive for fast access, but you can save a power and heat by going for a slower 5200RPM drive. If all it does is store near-line data and speed is of no concern, 5400RPM drives will be fine. It goes without saying to avoid 4200 RPM drives for your OS and applications at all cost.

Today, optical storage in a DVD is on the way out and can be omitted in your system specifications (I have not loaded a DVD in any of my computers in well over two years). Even so, It may be difficult to stuff more than one drive in a thin and light 15" model. Larger 17" laptops provide more options with regards to configuring multiple drives, and most come with two full 2.5" drive bays.

Mobile Features and Peripherals

Laptops are incredibly personal devices. When evaluating a laptop model, pay attention to the things that could make life easy or difficult. While it is highly beneficial to see, touch, and feel around a laptop before you purchase one, it's not always easy. Here are some key features and peripherals you want to have.

Docking Stations – Yes, You Need One

If you work on a laptop at a desk for any amount of time, don't get a laptop without a docking station, some dedicated monitors, and a keyboard. They make living with your laptop between work and home much easier, and you won't need to lug the power brick around.

Quality docking stations are purpose built for the laptop model they are intended to be used with. If you are buying the laptop from Dell, you want to get a Dell docking station. Aftermarket units can be a crapshoot. High quality laptops come with a special connector on the bottom for compatible docking stations, transmitting power, video, USB, and other external connections. If your laptop doesn't have a physical connector, you need to connect it with a special USB type cable which may not carry all of the desired signals for power, video, USB, and so on.

Ensure the docking station can drive two full size monitors at 1080p resolution (1920x1080). They may come with a combination of DVI, HDMI, DisplayPort, and VGA ports. I always recommend two DVI or DVI plus DisplayPort over having to use a VGA connection. Double check the video ports against your intended monitors before purchasing.

In my opinion, the best configuration I've found is to have at least two large 24"+ screens connected to the docking station. By having two large screens you can keep the laptop closed in the dock, perhaps tucked away under the desk. If you have only one external monitor, you can use the laptop screen for dual-screen use, but the screens will be of different sizes and resolutions, making it odd to work with.

Part of the pain of using a docking station is in how much room they take up. You have to physically affix the laptop to the dock and it needs to stay there, so if you are working on a small or cluttered desk the laptop + dock can take up a lot of room. I like to have mine on a small side table or on a rolling pedestal that slides under the desk.

New wireless docks stations are available which cut the cable clutter even more. A wireless dock looks more like a small brick with the cables out to your monitor and keyboard, but all you have to do is put the laptop somewhere near the dock to have it connect. They use the newer WiGig standard, so your laptop specification will need a WiGig card included to use it. However, they do have performance and capability restrictions which may be a deal killer, so review the specific unit carefully.

Ports

Laptops come with many ports and connectors to external devices. What you get and where they are located is an important topic to consider. For example, get as many USB 3.0 ports as possible, because you will use external drives and memory sticks more often than with a desktop. Look for USB ports which can power external devices like phones without the need for the laptop to be powered up.

Look for one or more USB-C ports. USB-C is a new connector standard that is poised to take over all other form factors for connectivity and power. Physically, It is similar in size to the Micro USB connector you might use on your phone, but unlike USB, and similar to Apple's Lightning connector, it is reversible, eliminating the bane of our collective existence for the past 20 years. Speaking of Apple and mobile computing, the new MacBook Pro models use only USB-C for all connections – power, video, networking, and peripherals.

However, USB-C is not a communications protocol like USB, VGA, or DVI. It is an emerging industry-standard connector for transmitting both data and power in one cable and can speak many protocols, not just one. It supports USB 3.1 (which, at 10Gb/s is 2x as fast as USB 3.0 and delivers much more power) as well as Thunderbolt 3. USB-C also supports DisplayPort, HDMI, and even VGA, allowing the same connector form factor to connect a wide variety of devices from video to power to external storage.

The present issue with USB-C is that there aren't a lot of devices using it at the moment, and there are millions of legacy devices out there which work just fine, so you may need to invest in one or more dongles to connect to removable drives and such. For a mobile user that is always a pain, but expect this to naturally go away over time as USB-C becomes the de facto standard across all devices.

Besides as least one USB-C port, you also want at least a full size HDMI and mini-DisplayPort to drive an external monitor without a docking station. Forget about VGA; they take up a lot of room and all newer screens come with at least one digital interface. Also, remember that Skylake drops support for VGA in its IGP. So if you are planning to use an older VGA-only monitor with a newer system, be prepared to get an adapter or, better yet, get a newer monitor.

RJ-45 rounds out the standard ports you need; do not settle for a work-capable laptop without a physical networking port. Otherwise you are hampering a fast machine with a slow wireless network connection.

As important as the ports are, you need them in the right place. If you aren't using a docking station you are probably going to plug in a mouse, power cord, and network cable to the unit itself. Are the ports out of the way of your hands when you are typing? I've had to use laptops where all of the ports were in exactly the wrong spot. Nothing was along the back, everything was along the sides closest to the front of the laptop, right where my hands naturally rest. Stupendously bad design.

Keyboard and Trackpad Options

The keyboard can make or break a laptop you use on a daily basis. This is subjective so I highly encourage you to sample the machine in a store somewhere, and extensively type on it. The feel and travel is what really separates good design from bad, and unfortunately manufactures will often replace an older unit with a great-feeling keyboard with newer model with a terrible keyboard (I'm looking at you, Lenovo). If you regularly use a docking station, the laptop keyboard may be of secondary concern until you take it on the road and try to type on it. Then you may be in for a rude surprise.

One of the laptop keyboard features I look for is a numeric keypad, usually found on higher-end, mobile workstation machines. This shifts the keyboard to the left a little and may take some time to get used to it, but I found it to be a real necessity. The position of the directional arrows and Insert/Delete buttons is also key to usability. Look for a dedicated Calculator button, it will gets a lot of use. It's a given that you will want media controls for volume and mute, so evaluate where they are as well. Note any special "Fn" key that works with the F-keys to provide additional controls. Is it in the right spot, or in the place where the CTRL key usually goes? If so, can you swap their location in the BIOS?

Laptops also have a trackpad, the flat area below the keyboard where you can emulate the mouse with your fingers. The mouse buttons (either physical or not) should be able to be assigned for left/right handed use in the driver. You may also have an eraser-like TrackPoint nub between the G and H keys to steer the mouse that you use in conjunction with the mouse buttons.

Like keyboards, trackpads are a personal preference and aren't all the same. If you use them regularly, any design-induced frustration can definitely be a deal-breaker. If the Trackpad is really large, you may inadvertently touch the trackpad when you are typing which moves the mouse and you end up with weird cursor behavior. The problem is you usually don't discover these design dangers until after you purchase and use the laptop for a while.

Communications

Get the most advanced wireless options available that support the widest array of standards, e.g. Dual Band, 802.11ac/a/b/g/n. It is only pennies more. Bluetooth 4.1 is a must; if you have a Smartphone or Bluetooth in your car it's a real benefit to transfer contact information or other near-field communications. It can also be used for wireless mice, negating the need for a USB transmitter.

If you are on the go without the ability to connect to Wi-Fi, you may want to invest in a 4G cellular modem. Optionally you may be able to tether your phone, turning it into a Wi-Fi hotspot, or invest in a personal Wi-Fi device from your cellular network provider.

Other Stuff You Want, You Just Don't Know It Yet

Speaking of keyboards, a backlit keyboard is a fantastic feature. It's also a cheap upgrade. It sounds dumb, but pay for it - you will not believe how handy it is when working at night.

Cameras and microphones are usually options you can specify to include or not, often in conjunction with the screen specification. Conspiracy types may want to eschew the camera for security reasons, as I would, although that's just because I look horrendous in a webcam. Built-in microphones are handy when conducting web-based meetings, although their sound quality is often poor.

If you plan on working a lot remotely without power, e.g. on airplanes, you may want to opt for a larger, 9-cell 91Wh (Watt Hour) battery. They are heavier to lug around all of the time, so I suggest a smaller 6-cell 72Wh battery instead which, with today's power-sipping processors, should last enough to get some things done. You can also opt for a 6-cell 91Wh "long life" battery but it is about $50 more expensive.

Regardless of whether you use a docking station or not, I also recommend a wireless mouse, preferably one with a very little USB transmitter that can stay inserted in the USB port without breaking off. I personally like the Logitech MX Master in large part because of the tiny transmitter. I've wrecked more than a few USB ports with longer dongles that get bumped around.

Get a good carrying case with pockets for stuff. Laptops require lots of extra doodads – power bricks, mice, portable drives, and so on. Backpacks are good, but the straps never last very long with heavier laptops. A soft, professional looking briefcase style is nice, and you won't look like a college student at client meetings. But if you have a huge power cord brick, make sure it fits without bulking out.

IX. Peripherals

The peripherals you outfit in your system may not be critical for overall performance, but they do affect day to day usage in important ways and thus should be a focus in specification.

Monitors

Monitors are, like hard drives, something that many people don't think twice about much when purchasing a new system, and that's rather unfortunate. They go for cheap and that is it. The choice of screen has a dramatic effect on your comfort and convenience level.

Today, large LED LCD screen are incredible bargains compared to their ancestors. It used to be that a good 21" CRT monitor with a resolution of 1280x1024 was about $1,900. Today's common 24" LCDs have almost double the resolution, use a third as much power, generate almost no heat, are uniformly bright with no distortion, and really good ones cost about a quarter of that to boot.

If one large LCD is great, two are better. The advantages of dual screen systems are hard to ignore. With the cost per monitor so low compared to the improvement in efficiency, it doesn't make any sense not to have two large screens at your disposal. Within a single application like Revit you can move the Project Browser and Properties Palette off to the other monitor, freeing up space.

For mundane tasks, the biggest boon multiple monitors have to productivity is in managing multiple applications, such as browser windows, email, office applications, Photoshop, etc. No one stays in only one program for very long; juggling multiple windows on one screen is maddening.

With 3ds Max Design, the use of three monitors (which may require >1 graphics card) is entirely warranted. That way you can have the application window on one screen, the scene explorer and material editor on another, and a curve editor on the right, all at the same time – that's efficiency.

With multiple monitors comes the need to properly manage the windows. The ability to spread an application across both screens or to bounce a maximized window around is highly productive. To that end I recommend investing in a multi-monitor utility such as UltraMon or DisplayFusion. Each offers the ability to put additional buttons in the top right corner to manage windows across screens. If you have an Nvidia card, the driver includes an "NView" utility which can offer up some similar functionality as well.

Monitor Technology

There are differentiators between monitor models which do make a difference. First and foremost is technology used for the panel itself. There are three main technologies: TN (Twisted Nematic), VA (Vertical Alignment), and IPS (In-Plane Switching). Of the three, look for and demand an IPS panel. At every metric, IPS panels deliver sharper, almost reference like quality.

Technology	Color Reproduction	Viewing Angle	Response Time	Price	Comments
In Plane Switching (IPS)	Excellent	Excellent	Good	Expensive	Slight color tinges may be visible at an angle
Vertical Alignment (VA)	Good	Good	Average	Reasonable	Colors shift when viewed at an angle
Twisted Nematic (TN)	Average	Average	Excellent	Inexpensive	Limited to 6-bit color; Technology improving.

There are several iterations of IPS panels, such as S-IPS, H-IPS, e-IPS, and P-IPS. They are all relatively similar to each other and not really worth getting into detail. Until the holy grail of flexible OLED monitors are commercially viable, an IPS screen is what you want.

The second concern is size and screen resolution. Probably the most practical size in a dual screen configuration is a 24" LCD, although 27" screens are becoming more popular as prices drop. Good 30" LCDs are still considered a luxury as their prices are over $1,000. However, if you can swing one (or two), they are definitely worth it. Monitors will last several machine generations, so it pays to remember it will be around for a while so buy the largest one you can.

While TN is inexpensive it did have a reputation for being cheap but with poor to average color fidelity and overall quality. That's changing a little as TN is improving around the edges; look at specific reviews before jumping on a TN panel.

4K Displays

The resolution (number of pixels horizontally and vertically) determines how much of your application and user interface is available to you at one time. Currently the Holy Grail of hype today are so-called "4K" displays, which is often and incorrectly used to refer to the more common standard UHD, which is 3,840 x 2,160. That's 4x as many pixels as a typical 1080p 1920x1080 monitor.

With so many pixels, your 4K monitor will need to be at least 24" or larger. Graphics cards need to be good enough to spit out that kind of resolution, but today's cards do not have much of a problem on that end. They can and do look fantastic, but as discussed in the section on laptops, the higher resolution can have drawbacks, in terms of application support and scalability.

The bottom line on 4K resolution is that while display technology is improving to incorporate 4K as a new standard, it has some shakedown cruises to complete. You need to check all of the specifications carefully as well as your applications and ensure things match your graphics card's capabilities. Remember too that lower resolution 1080p displays are just fine for anything you need to do in the Building Design Suite and AEC Industry Collection.

Resolutions You Want

Because UI elements are metered in pixels, the more pixels you have on screen the more stuff you see at one time, and the more you can get done without fiddling with pans or scrolling windows. That time all adds up. With the Ribbon being so prevalent in Revit, smaller resolutions below 1080p chop the icons down to smaller bits making the interface slightly harder to use; you may also end up collapsing it down because of the lacking vertical dimension.

For a 24" monitor, look for a resolution of no less than 1920x1200. Most cheap 24" screens max out at 1080p, or 1920x1080, which is acceptable to most but you lose over a hundred pixels in height. Furthermore, those wide aspect ratio screens aren't the most productive for Revit, 3ds Max and other Autodesk apps, which appreciates the height to handle the Project Browser, property palettes, the command panel, and other vertical UI elements.

Other niceties on good LCD panels are USB ports and memory card readers, which put things closer to your desk rather than being in your PC, which might be on the floor or otherwise hard to reach.

I've seen people using wall mounted 48" LCD televisions with success. You just need to ensure you can run a completely digital signal out to the TV. You'll probably need to invest in a DVI-D->HDMI adapter if you do not have an HDMI Out port on your machine. Don't fall back to VGA. Remember that TVs will probably max out at 1280x1080 resolution, which is smaller than you can get on good quality 24" and larger monitors. Because they are physically larger, the pixels will also be larger, so ensure you are far enough away from the TV that the coarseness doesn't become an issue.

HDMI and DisplayPort

HDMI (High Definition Multimedia Interface) is a standard digital video interface found on modern televisions for carrying audio and video from cable boxes and other digital A/V equipment. It also works very well for computer monitor signals. Usually found on less expensive, consumer-oriented monitors, it is nonetheless a proper replacement for analog VGA, which by now should be avoided. HDMI is nice because there is a simple push-in connector; there are no little thumbscrews to deal with as with VGA and DVI, making it much easier to connect and disconnect your cables.

Another digital interface is DisplayPort. This is similar to HDMI ports but slightly bigger and are often found on more high-end monitors and graphics cards. Like HDMI, DisplayPort is an interface meant to replace VGA and DVI, and can be used to carry audio, USB, and other forms of data. It packets the data, much like Ethernet and PCI Express. The current standard is DisplayPort 1.4, published in March of 2016, and most late model graphics cards and monitors support it. DisplayPort itself is a drop-dead simple form factor to use, and in my opinion superior to HDMI because it has a locking mechanism so it cannot come out accidentally. Just push and click, no knobs or screwdrivers needed.

Mice and Keyboards

Along with the monitor the most important thing to choose wisely is your mouse and keyboard. Don't settle for the cheap plastic things given out by your system vendor. They almost all universally stink while the options out there are truly great. You are going to handle each of these every day for years, so it pays to find ones you like. That said, I'm going to provide my personal recommendations.

For both keyboards and mice, I highly recommend wireless devices. They can be integrated as one solution with some manufacturers. They can either be radio controlled (RF) or implemented via Bluetooth. Logitech and others offer Unifying Receiver technology that allows more than one compatible device to be hooked into a single USB receiver, so you can mix and match their wireless products pretty easily.

Personal Recommendations:

Mouse: Logitech MX Master. I graduated from destroying four or five Logitech Performance MX mice over the years to this one, and it's better in almost every way. For me the killer function is the thumb button, which is molded into the mouse body, not a separate physical button, so you aren't tripping over it.

Through the driver, I configure it as a Middle Button. This allows me to pan/orbit in applications using my thumb instead of moving my index finger from the left button to the wheel. Because of this, I can select an object in Revit, start to drag it by holding down the left button, and while still holding the LMB down, use my thumb to pan the screen at the same time. It's an incredibly efficient way to drive the software and quite unwieldy to do (for me anyway) with the middle mouse button and left mouse buttons alone.

Keyboard:

Keyboards, like mice, are very personal devices. As a primary instrument to interfacing with your computer, spending time with a good one will make your daily life much easier. Typing for extended periods of time on bad ones can be debilitating. If you play games you need one that is ready to take on the abuse.

In my view a good keyboard offers the following things: excellent key feel, a separate numeric keypad, and dedicated multimedia and mini-application controls. As with mobile keyboards, I loot for a dedicated Calculator button. Other may appreciate remappable keys which can fire off specific commands. I also highly appreciate backlighting for working in low light.

Feel is purely subjective, so it pays to shop at a brick and mortar store and play with the keys for a while. Manufacturers have clued into this aspect of the enthusiast market, and built high-end mechanical keyboards for discriminating users. Some of you may remember the old IBM keyboards, which were built

like tanks, weighed about as much as one, and had a very heavy feel to the key switches giving off very audible clicks when depressed. This went away with the introduction of cheap keyboards and laptop chicklet keys. Now manufacturers are bringing heavy back with the use of Cherry MX switches.

Cherry Corporation, founded in 1953, is actually the oldest keyboard manufacturer in the world. They produce a line of keyboard switches, the Cherry MX series, which come in several color-coded varieties that vary in actuation force (measured in centi-Newtons, or cN), tactile feel, and audible click noise.

Linear switches are the simplest type, moving straight up and down without any additional tactile feedback or loud clicking noise. In this lineup, the Cherry MX Black switches have a medium to high actuation force (60 cN), making them the stiffest of the lineup. They are most often used in point-of-sale equipment. Cherry MX Red switches have a low actuation force at 45cN, marketed for gamers where rapid actuation is important.

Other types of switches add tactile feedback and varying levels of click noise. Cherry MX Brown switches are a popular switch with a tactile but non-clicky action. It provides excellent feedback without the noise associated with most tactile keyboards. Cherry MX Blue switches are tactile and clicky, favored by heavy typists due to their tactile “bump” and click sound, but do have a higher actuation force (50 cN).

X. Operating Systems – Windows 10

As far as choice of operating system goes, luckily this is a pretty easy decision. Your choices are narrowed down to just three. You could go with either Windows 7 or Windows 8.1, but in the end you'll probably end up with Windows 10 on the hard drive - whether you want it or not.

Regardless, it's quite time enough for most users to forget Windows 7/8.1 and move to Windows 10. Some of the new features and benefits to the user include:

- Last year, a Windows 10 upgrade was free for Windows 7/8 users. No license key was required – it picked it up from your Win7/8 installation. Alas, the promotion is over and the upgrade costs $119.
- When either upgrading an existing Windows installation or performing a clean install, Windows 10 is very easy to download and simple and fast to install. Using Windows 10 Media Creation Tool, you can easily drop it onto a USB drive for installation.
- Although not officially supported, I have ***personally*** seen no incompatibilities with anything from Autodesk, at least as far back as 2014 releases. That is not to say they don't exist, but with patches and the proper Windows Updates installed they should be minimal or fixed.
- Consistent interface regardless of device type. Tablets operate just like desktops with a touch UI.
- DirectX 12 should bring a 20-40% increase over DirectX 11 in applications that make heavy use of the API, such as those in the Building Design Suite. Windows 7/8 will never have access to DX 12.
- Solid performance all around with support for more devices and newer technologies such as NVMe SSDs. All of my computers' devices just seemed to work, from laptops to desktops.
- Virtual Desktops allow single-monitor users to access multiple virtual desktops handy for splitting up usage between work and play, or different projects, or whatever.
- System requirements are largely the same as Windows 7/8, but with better overall hardware support.
- Redesigned Start Menu, a combination of Windows 8 Live Tiles and Windows 7 functionality. No more weird charms bar as in Windows 8.1. All settings and applications are easily accessed in the Start Menu.
- Enhanced Store provides access to music, videos, games, and apps. Apps from the Store will run on every device – PC, laptop, tablet, or phone.
- Cortana, a personal assistant, much like Siri.
- Unprecedented (and potentially problematic) new policies to enforce updates via Windows Update when you may not want them. However, you will get security patches applied right away which should improve security inside the firewall. Expect large rollup updates every few weeks.
- Microsoft Edge browser - very fast, very lean, boatloads of new features.
- All "apps" run as foreground windowed applications.
- Windows Snap – snap windows to one half of your screen. Works great on dual monitor systems. Use Windows Key + arrows to move windows around the screen.

If you already have Windows 10, make sure you get the new massive Windows "Anniversary Update" [33]which rolled out in August of this year. Some of the new features:

- Redesigned Start Menu tweaks provide a sleeker and more intuitive interface to access system settings.
- Improvements to Windows Ink for tablet folks
- Fantastic scenic Lock Screen wallpapers update regularly, and new media controls allow playback without unlocking the PC.
- Cortana updates to make her (it) more helpful
- Windows Defender improvements

[33] http://www.howtogeek.com/248177/whats-new-in-windows-10s-anniversary-update/

XI. Build or Buy

For most people who specify BIM and 3D workstations for their company, the question of buying a packaged workstation from a system vendor like HP, Lenovo, BOXX, or Dell is a no-brainer. For others they may want to build it out themselves. The question comes down to cost per unit performance, availability of specific components, warranties, and servicing levels. This section provides some insight as to the pros and cons of building verses buying complete systems.

Buy it!

For many people, particularly those in business environments, purchasing an assembled machine is clearly the way to go. You probably already have a preferred hardware vendor, internal IT support, and a host of corporate IT policies and bylaws, so building an out-of-spec workstation which is not blessed by your IT folks or backed up by a system-wide warranty is considered risky and a non-starter.

The Pros:

Buying a workstation from a vendor is largely about minimizing risk and maximizing uptime. The onus is on the vendor to test all of the components, install the OS and the drivers, and make sure it works before handing it off to you. All you should have to do is unpack it, turn it on, and install your software.

Vendors design and build professional-grade workstations as complete systems that can be tested and certified by Autodesk and other software vendors. They specify components which are manufactured under stricter quality control than ones in typical desktops.

By establishing a solid business to business relationship with your vendor, they will work with you to configure systems more to your liking and provide benefits such as free shipping and notifications of sales and special offers.

Warranties, especially next day repair services, are crucial for keeping things humming along and are highly recommended. If you laptop's video card goes south you can't just pop down to best Buy to pick another one up. Having a certified mechanic come to your place of business and replace the card on the next day justifies the warranty.

The Cons:

The primary issues with packaged workstations are:

- Typically meh to lousy choices in preconfigured models, especially if you know what you want.
- Inflexibility in customizing your system configuration.
- Unreal upcharges for trivial upgrades or changes from preconfigured models.
- Proprietary components.

System vendors love creating packaged system configurations, building a bunch of them, and shipping them on demand. When you have customers that need 20 workstations that are already sitting in the warehouse, that's easy to deal with. When you have to build 20 custom machines that have non-standard parts, shipping will be delayed and errors and omissions can be more frequent. But that's part of what they do on a daily basis.

When a vendor like Dell or HP designs a line of workstations, they configure several models that align with certain price points and requirements. The basis of each new model has a range of available processors, RAM configurations, graphics cards, and hard drive sizes and configurations. The components are easy for them to get and provide a decent range of performance and price points.

The problem is that typically their pre-packaged systems have non-optimal configurations. They default to the lowest-end / last year's CPUs, incorrect amounts of RAM, terrible video cards, and slow hard disks, solely to make the advertised price artificially low. Once you start upgrading components to something that fits your needs, the price quickly skyrockets.

You would often have ridiculous upcharges to go from 8GB to 16GB or swap out a 1TB hard drive for a 512GB SSD. They may entice you with a low prices of $999 but after you configure it for real-world use it's like $2,999. This isn't so bad today, as the preconfigured systems I see usually have 16GB of RAM and an SSD, but component upcharges are still a hard pill to swallow, especially when you see the Newegg price.

Vendors like Dell and HP custom design their cases, motherboards, and power supplies, all meant to provide adequate reliability for the lowest price to the vendor. Above all, the system must be stable, so you won't find many knobs to tweak in the system BIOS that you would in an enthusiast motherboard from ASUS, for example. The power supply won't have the nice fan controls or modular cables or gold-encrusted capacitors and such. The memory modules won't be painted neon pink with little finger-like heat sinks on them. In short the system components are meant to do a job, not look snazzy doing it.

One pretty popular vendor who caters to the design world is BOXX, who can configure systems to your exact specifications. Except for the case, you will find these are largely built from stock, off-the-shelf parts you could put together yourself for much less.

The ability for you to swap out items is limited, depending on vendor. Remember that the completed system needs to be certifiable, and Autodesk won't certify systems with "gaming" graphics cards, even if they run rings around the professional cards. Whether this is a conspiracy between Autodesk and Nvidia/AMD to artificially keep their workstation lines viable is up for debate. For the base workstation platform that's not much of an issue - vendors will usually make available all appropriate CPUs for that particular machine. Traditional vendors like Dell nor HP will not swap out a Quadro M2000 for a GeForce GTX 1080 because it can't get certified. Others like BOXX will.

Tips and Tricks for Purchasing Packaged Systems

When purchasing a machine from an established vendor, it helps to understand some tricks of the trade.

1. **Always talk to a real person and establish a professional corporate relationship with the vendor.** In the cases of Dell and HP, you will likely make use of their online stores to do research and initial pricing, and perhaps save a cart of your prospective machine. Look to make them a partner. You will always save money by doing so. The online store is designed to allow people easy ways to purchase machines, but more components and options are available if you have an account and go through your salesperson. Try to establish the highest tier of service that you can. For example, Dell has a "premier" customer status that provides decent service and a good online support portal for tracking your machines, their warranties, and service histories. Use these things to your benefit. Good corporate customers often get freebies such as free shipping or deals on specific components.
2. **Once you narrow down your system, do not pay more for any upgrade than absolutely necessary.** Disregard the online upcharges you see on web site stores. They are fictional if you are working through a sales rep. For example, you will typically find Dell and HP have exorbitantly high RAM upgrade prices. Price out RAM from someplace like Crucial.com, which has the exact same module specification for your particular workstation. See if your salesperson can negotiate the RAM upgrade charge. If they cannot match Crucial.com's price, buy the system with the least amount of RAM, and purchase the RAM upgrade elsewhere. This is one time where installing the RAM yourself can save a lot of money, particularly if it is more than one system. RAM is RAM - Dell doesn't sprinkle magic dust on it that makes it work better with their systems. They get it from the same place you would, so don't pay more for it.

3. **Get the physical media for all of the software you are purchasing**. At a minimum this would include the OS, but also Office and the driver disks as well. I once purchased a number of systems for a customer to find that one of them shipped without anything installed on the hard disk. No OS, nothing. It was completely blank, so having the resource DVDs on hand was a godsend.
4. **When you receive a quote from your vendor, check it over – twice**. It's easy to for them to key in a slightly wrong thing. The wrong CPU, video card, RAM configuration – you name it, it can get screwed up. Out of a hundred quotes that I have received over the years, easily 40% of them had some sort of tiny mistake, such as extra shipping costs, wrong RAM configuration, and so on.
5. **Evaluate and formally approve the extended warranties that can and will be tacked on to the machine.** However, vendors will often slip an extended warranty (with an extended price tag) into your system quote and hope you will overlook it. Depending on your IT policies and expected lifespan of the machine, this may or may not be something to consider. When disaster strikes it's nice to have their servicing agent come onsite and replace the part for free. Generally speaking the 3 year onsite warranty that comes with most vendor systems is worth it for the corporate machine. It's usually inexpensive and pays for itself if the machine blows a fuse. After three years, consider the warranty cost carefully against the expected replacement time for the machine; the possible exception to this is with laptops, where I recommend you carry over warranties after the original one expires.

Build It Yourself!

On the other hand, you can build your system from scratch. If you have done your research and are looking to ensure your system has specific components, and you don't mind the labor in putting together your components, building your new workstation makes sense. I know many small companies that do this regularly instead of going to a name vendor. Just be aware of the limitations of a BIY approach.

The Pros:

The more you know about specific components and your specific workstation needs, the more a BIY approach makes sense. As we've seen in the graphics benchmarks, the best card for a particular job is possibly not on the model's available list. If a system vendor doesn't provide a particular option you like, there is little reason to pay money for what you do not.

If you are an overclocker obsessed with squeezing out every bit of performance potential, the BIY route is definitely the way to go, as you can outfit it with the right motherboard and exotic cooling options to ramp up the CPU and memory timings. "Enthusiast" motherboards have less conservative timings and solid quality capacitors to enable higher performance levels.

If you are looking for alternative builds, such as small form factor (SFF) cases or even building your own custom case - and people do this quite a bit - then obviously you are looking at a BIY approach. Even with an off the shelf case, it gives you the opportunity to physically put it all together and optimize where everything is, so you can route the cables or place the drives in the best location to afford the best airflow.

The Cons:

Building systems comes with its own issues and risks. Researching the parts usually takes longer than picking from a no-name list on a configuration page. You also have to pick the motherboard, RAM, case, and power supply, all of which would come by default in a vendor's machine and all of which have tons of candidates to choose from.

Second, you can't be sure the parts you get will actually work when you get them assembled. Defective parts could need to be RMA'd back to the store or manufacturer, delaying when you can start using the system. Components could fail early after the initial build and need to be sent back. Bad batches of components can make this happen more than once - it can seriously be a real pain.

Third, true workstation components such as Xeon CPUs and motherboards that support them are not commodity parts and thus are not easily available or inexpensive. Some components are only sold to system integrators, so there is a ceiling as to what you can build yourself.

Lastly, vendor workstations come with at least three year system-wide warranties, and repair parts can be overnighted to you. You usually get one-year warranties on most parts (drives may be up to 5 years) but won't get the same quick turnaround with a bad part from an online retailer, increasing downtime.

Tricks for Buying Build It Yourself Systems

1. Most online retailers have good RMA programs in case something doesn't work, so seek out reputable companies with good return policies.
2. If you are able to get everything from one place, you can save on overall total shipping costs.
3. Most online stores have "wish lists" and shopping carts, so configure a few of these for research. Check it every so often as prices fluctuate day to day. If you see a nice rebate happening at the moment, strike quickly because they will be gone at some point.
4. When researching components, pay strict attention to the reviews for each one. Often it is the little things that turn a great product into a problem due to combinations of factors. Look at components with many reviews, as this points to their popularity and should provide a better quality evaluation.
5. Make sure you have enough cables. SATA hard drives come either as a full kit with a cable, or just the bare drive. Motherboards will typically come with enough cables and brackets for most builds, but you may need extras.
6. Get a package of small zip-ties to tuck away the power and drive cables to make things neat and improve air flow. Use a small "flush cutter" tool which easily and cleanly cuts zip-ties.
7. Check the specific processor for the inclusion of the heat sink and fan: Broadwell E and Xeon boxed CPUs typically do not come with a cooler. Air coolers will work within the processors temperature envelope, or you may opt for a different cooling solution altogether like a closed loop water cooling system and radiator to take the temperatures even lower. This is critical for overclockers and people who run their systems at higher levels for extended periods of time.
8. Lastly, give yourself time to put it together right. A basic system build, from the unboxing of parts to turning on the power button is still a few hours. It's not rocket science but a little bit of planning helps.

Hybrid Builds

It is possible you may get the best of both worlds by buying a bare-bones workstation from your vendor to get the warranty and base build quality, and purchasing the stuff you really want separately, such as video cards, SSDs, and peripherals.

A vendor's base workstation machine will come with a motherboard, case, CPU, a hard drive, and at least some memory, and a chintzy graphics card. Get the fastest CPU you can afford with the least RAM, the smallest mechanical storage subsystem, and just built-in graphics if available, since those are the parts you want to be choosy about. Then purchase the components you really want elsewhere. You may initially be paying a little extra for extra parts, but keep them on hand as spares in case anything goes south.

Tips for Setting Up New Systems

Whether you buy a system or build it from scratch, there is a standard series of startup tasks I recommend to make sure everything is in working order.

1. Once the machine is initialized and you get to the desktop for the very first time, *confirm all components and benchmark the machine in CPU, memory, and graphics* using standard free benchmarks such as SiSoft Sandra 2015, 3D Mark11, PCMark8, Cinebench, and others. Anything you do to the machine from here on out - adding software / hardware, tweaking registry settings, running applications, and so on - will change your machine's performance profile.
2. Once initially benchmarked, *immediately check the machine for "craplets" and uninstall them.* Depending on the system this could be included backup software, update utilities, printing utilities, or picture management apps which are worthless. Go to Task Manager's Startup tab in Windows 8/10 to see what applications are starting up with the system and disable them. If you know your way around Windows Services, optimize the machine further by disabling application specific services that do nothing but consume memory space and CPU cycles. Benchmark the system again and see if removing unnecessary apps helps things out.
3. *Install and run the latest version of CCleaner and perform a cleanup of the Registry*. Also root out unnecessary startup items and uninstall anything that looks sketchy. In Windows 8.1/10 uninstall any tablet-like "apps" you do not need, which just chews up valuable SSD space.
4. Disable Windows 10's propensity to install hardware drivers. Hardware drivers provided by Windows 10 are usually stripped down versions and may not work properly with certain Autodesk applications. Research and install those yourself. The problem is that Windows 10 will automatically look for and install drivers. There is a "Never install drivers from Windows Update" option in the latest version of Windows 10, buried in the old Control Panel. Go to System > Advanced Settings > Hardware Tab, Device Installation settings, and select No.
5. *Run Windows Update*. Every system I've ever received was outdated out of the box. Make sure all of your patches are applied before starting any application installations. Ensure you are not installing any Microsoft-recommended drivers at this point (see tip #5).
6. *Update the drivers from the hardware manufacturer*. This is something a surprising number of people do not do, particularly with new workstations. Immediately check the component manufacturer's or system vendor's support & driver site for motherboard-specific driver updates for the BIOS, chipset, audio, USB, and network. Go directly to Nvidia or AMD for video card drivers. Dell and HP typically update 3rd party board drivers once in a blue moon, whereas Nvidia and AMD will have new drivers every month. Just updating drivers can have a significant effect on performance, so benchmark again.
7. Finally, install your applications and tweak any post-install settings. For Autodesk installations, always clean out your %TEMP% files and reboot the system before installing. After installation, launch each application and accept the EULA agreement. Launch the Autodesk Desktop App (formerly the Autodesk Application Manager) to see how many updates are available. With a clean install of the applications within the Building Design Suite Premium 2017, expect to spend about an hour simply downloading and installing service packs and updates.

Once all of your apps are installed, again check what application specific startup processes or running services are unnecessary. Lastly, right before you really start using it, benchmark it again.

XII. Matt's Workstation Buying Guide, 2016 Edition

In this section I originally intended to build three classes of machines, based on the three user types we identified earlier:

1. A Grunt machine, based on the Skylake desktop platform;
2. A BIM Champ machine, based on the Intel HEDT (High End Desktop) Broadwell E platform;
3. A Viz Wiz workstation, based on the Xeon E5-26xx platform.

Build Philosophy and Methodology

For each build class I was going to strive to specify two machines: A Dell Precision Workstation and a BIY Built It Yourself) system with components priced from Newegg.com. We would then compare prices and declare a winner. I selected the Dell Precision for two reasons: First, they are very popular with the design world and many people are familiar with their systems and have accounts with them. Secondly, they have configurations which are similar for other similar top-tier vendors like HP and Lenovo.

For the BIY systems I use Newegg.com because they have aggressive pricing and can be a single source for all parts. This should save on shipping costs and general pain, because if anything goes wrong there is one place to go. Newegg.com has a pretty good reputation for low prices and low hassles if something needs to be RMA'd.

However, it becomes apparent that while it's easy to find desktop and HEDT based parts and build out those two workstations, it's much more difficult to build out a Xeon E5-based workstation that would be considered cost effective from parts. For that I simply priced out a single hybrid build with a Dell Precision Workstation along with upgrade components from Newegg.

As mentioned earlier, specifying a workstation from a vendor is full of compromises, which I hate. They usually don't offer the right graphics card, so you either buy a lower-end one or a more expensive "workstation" class card that performs the same or worse than the one you wanted.

Due to these limitation I decided to build the Dell workstations as hybrid builds, purchasing as much from Dell as I could and supplementing it with the right parts from Newegg.com. In other words, I did not let my specification become compromised because Dell did not allow me to purchase a particular part.

Base Specifications and Terms:

- I'm pricing everything anonymously from online sites. For Dell, if I requested actual quotes from human beings, even salespeople, the price would probably be lower.
- I'm including dual Dell UltraSharp 27 InfinityEdge U2717D IPS flat panel monitors in all builds. At $420 from Newegg, it's a fantastic deal and gets very good reviews.
- No MS Office, productivity software, security software, or specific vendor tweaks are included.
- For the Dell systems all DVD resource disks for the OS and drivers are included in the price.
- Shipping prices are not included on any builds.
- Newegg prices include any current as-of-this-writing rebates and special offers.
- For the video card, I'm specifying as a standard the GeForce GTX 1070, which provides excellent performance for the money.
- Unless forced to do so, I did not specify optical storage. Who uses DVDs anymore?

The Grunt: Dell Precision T3620

First up is a Dell Precision T3620 Workstation in a mini tower form factor.

The Grunt Workstation: Dell T3620 Precision Workstation, Mini Tower		
Component	Item	Price
Processor	Intel Xeon E3-1270 v5 Quad-Core @ 3.6GHz	$2,013.34
Memory	32GB (2x16GB) DDR4 2133 Non-ECC	
Graphics	Nvidia Quadro NVS 315, 1GB	
Disk	M.2 512GB PCIe NVMe Class 40 Solid State Drive	
Power supply	365W Up to 92% Efficient Power Supply	
Keyboard	Dell KB216 Wired Keyboard Black	
Mouse	None (see below)	
Resource Disk	Windows 10 OS Recovery and Resource DVDs	
OS	Windows 10 Professional, 64-bit	
Warranty	3 Year Hardware Service with Onsite/In-Home Service after Remote Diagnosis	
Dell Subtotal		**$2,013.34**
Additional items purchased separately from Newegg.com		
Video Card	EVGA GeForce GTX 1070 w/8GB	$399.99
Monitors	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
Mouse	Logitech MX Master Mouse	$69.99
Newegg Subtotal		**$1,329.96**
System Total		**$3,343.30**

Notes

1. Dell provides this model with either a Xeon E3-1270 v5, a Core i7-6700, and a Core i7-6700K processor for just around the same price (about an $80 total difference between the three).
2. The Xeon does not have an IGP so a default video card must be purchased.
3. The 512GB M.2 PCIe NVMe drive chosen here will be the standard drive specified for all builds.
4. The U2717D monitor is available from Dell at $569.99, but not available as a discounted upgrade component. It's only $429.99 from Newegg, for a total savings of $280.
5. The 3620 has no advanced USB 3.1 or Thunderbolt ports on the system.

The Grunt: Newegg Edition

Next up is a comparable home-brewed BIY system completely from Newegg:

The Grunt: Newegg		
Component	**Item**	**Price**
Case	Fractal Design Define R4 Black Silent Mid-Tower Computer Case	$109.99
Processor	Intel Core i7-6700K Skylake 4.0GHz LGA 1151 Processor	$339.99
Motherboard	ASUS Z170-P LGA 1151 motherboard	$109.99
Memory	Corsair Vengeance LPX 32GB (2 x 16GB) 288-Pin DDR4 SDRAM DDR4 2133 (PC4 17000) Desktop Memory	$167.99
Graphics	EVGA GeForce GTX 1070 w/8GB	$399.99
Storage	Samsung 960 Pro M.2 512GB Solid State Drive	$329.99
Power supply	Corsair AX Series AX 860 860W 80+ Platinum PSU	$169.99
Mouse	Logitech MX Master Mouse	$69.99
Keyboard	Logitech G610 Orion Red Mechanical Gaming Keyboard	$99.99
Monitors	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
OS	Windows 10 Professional, 64-bit, OEM	$139.99
System Total		**$2,797.88**

Notes

1. Because motherboards for the Xeon E3 are difficult to get on the open market, this system was specified with a common Skylake i7-6700K processor @ 4GHz with a Z170 motherboard.
2. The ASUS Z190-P motherboard is a good basic selection with little frills but good performance.
3. The Samsung 960 Pro M.2 SSD is regarded as top of its class in benchmarks, but not widely available at the time of this writing.
4. I selected a “gaming” keyboard with Cherry MX switches with a budget of $100.
5. 32GB as 2x16GB modules allows for another 32GB upgrade in the future.

The Grunt BIM Workstation: Build Analysis

I would not use the term “entry level” for this workstation; it’s a professional-class build meant for people using everything in the AEC Industry Collection, and will handle the vast majority of the tasks with aplomb.

Overall the Newegg.com system was about $545 less than the Dell for the same basic configuration. However, the Newegg machine has a faster CPU (4GHz vs 3.6GHz), a better keyboard, and a known performer for the SSD. The Fractal Design case is serviceable but not as high quality as the Dell Precision mid-tower; assume the Dell system to be nearly silent but the Newegg system to emit some fan noise.

The Dell was more expensive primarily due to the 32GB of RAM, which was a staggering $494.79 more than specifying it as a single 8GB stick, which is $326.80 more than the entire 32GB kit from Newegg.

As I wasn’t willing to accept the “workstation” class video cards Dell offered, I specified a cheap Quadro NVS card and picked up the GTX 1070 from Newegg. I could have saved this cost ($79.37) by specifying the i7-6700K CPU, which has an IGP, but the upcharge for the i7-6700K was $53.74, negating any real benefit. And I would be stuck with an IGP to handle in addition to a discrete graphics card.

The downsides are what you would expect of BIY systems: No on-site service warranties, higher propensity to RMA back a defective component, and the time it takes to build it and install the OS.

The BIM Champ 6-Core Xeon Workstation: Dell Precision 5810

Stepping up to 6 CPU cores, we have the following Dell Precision 5810 based on the Xeon E5-1650 v4. This Xeon is very closely aligned with the Broadwell E i7-6850K which we can use for the BIY build.

The BIM Champ 6-Core Xeon: Dell Precision Tower 5810		
Component	Item	Price
Processor	Intel Xeon E5-1650 v4 6-Core @ 3.6GHz	$3,506.39
Memory	32GB (4x8GB) DDR4 2400 RDIMM ECC	
Graphics	Nvidia Quadro M4000 8GB card	
Storage	512GB 2.5" SATA Class 20 Solid State Drive	
Keyboard	Dell KB-216 Wired USB Keyboard Black	
OS	Windows 8.1 Professional, 64-bit, w/DVD Recovery	
Warranty	3 Year ProSupport with Next Business Day Onsite Service	
Chassis Option	Dell Precision Tower 5810 825W TPM, BW	
Resource Disk	Windows 10 64-Bit OS Recovery and Resource DVDs	
Dell Subtotal		**$3,506.39**
Additional items purchased separately from Newegg.com		
Video Card	EVGA GeForce GTX 1070 w/8GB	$399.99
Monitors	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
Mouse	Logitech MX Master Mouse	$69.99
Newegg Subtotal		**$1,329.96**
System Total		**$4,836.35**

Notes

1. Dell's online configurator does not allow for a 512GB M.2 boot drive, so this system is being specified with a traditional 512GB SATA SSD.
2. As with the Grunt Precision build, you must specific a video card. However, the lowest you can go on the Dell 5810 is a dual 4GB AMD FirePro W5100 graphics card. Because we want to stay with Nvidia for comparison purposes, we will upgrade to a Quadro M4000 8GB card. The M4000 has 1664 CUDA cores and 8GB of RAM, which is somewhat comparable to the GTX 1070 with 1920 CUDA cores and 8GB of RAM.

The BIM Champ 6-Core Broadwell E : Newegg

The BIM Champ 6-Core Broadwell E: Newegg		
Component	**Item**	**Price**
Case	Fractal Design Define R4 Black Silent Mid-Tower Computer Case	$109.99
Processor	Intel Core i7-6850K Broadwell E 6-Core 3.6 GHz	$609.99
Processor Cooler	Corsair Hydro H100i GTX Extreme Performance Water/Liquid CPU Cooler	$104.99
Motherboard	ASUS X99-E LGA 2011-v3 Motherboard	$209.00
Memory	Corsair Vengeance LPX 32GB (2 x 16GB) 288-Pin DDR4 SDRAM DDR4 2133 (PC4 17000) Desktop Memory	$167.99
Graphics	EVGA GeForce GTX 1070 w/8GB	$399.99
Storage	Samsung 960 Pro M.2 512GB Solid State Drive	$329.99
Power supply	Corsair AX Series AX 860 860W 80+ Platinum PSU	$169.99
Mouse	Logitech MX Master Wireless Mouse	$69.99
Keyboard	Logitech G610 Orion Red Mechanical Gaming Keyboard	$99.99
OS	Windows 10 Professional, 64-bit, OEM	$139.99
Monitor	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
System Total		**$3,271.88**

Notes

1. Because the Broadwell E CPU does not come with a cooler, a Corsair water cooling system was added.
2. The ASUS X99-A is a no-frills but competent motherboard for this build.
3. The GTX 1070 was specified to maximize value. The Quadro M4000 retails for $766.99 and is built on the previous generation GM204. It has slightly fewer CUDA cores than the GTX 1070 but has the same 8GB of video RAM.

The BIM Champ Build Analysis

For this system I decided to step things up to the next logical level; a 6-core Haswell-based Xeon E5-1650 v4 CPU in the Dell Precision 5810 and a Broadwell E i7-6830K in the Newegg build. Comparing the CPUs, the Xeon E5-1650 v5 has a Passmark score of benchmarks at 14,372. The Core i7-6850K has a score of 14,318, making them for all practical purposes identical given the vagaries of Passmark scores.

However, as usual, the issues with specifying Dell workstation components brings on the headaches. Because the list of video cards is limited to professional level cards, I had to specify a Quadro M4000 for the Precision to equate to the graphics card of choice in the BIY systems, the GTX 1070.

As with the Grunt build, the BIY system wins in the price / performance category. It is over $1,500 less than the Dell Precision but has more graphics power, better cooling, a much better hard disk, and a better keyboard.

The Viz Wiz 8-Core Xeon Workstation : Dell 7810

The Viz Wiz 8-Core Xeon: Dell Precision Tower 7810		
Component	Item	Price
Processor	Intel Xeon E5-1660 v4 8-Core @ 3.2GHz	$5,880.72
Memory	64GB (4x16GB) DDR4 2400 RDIMM ECC	
Graphics	Nvidia Quadro M4000 8GB card	
	Nvidia Quadro M4000 8GB card	
Storage	512GB 2.5" SATA Class 20 Solid State Drive	
Keyboard	Dell KB-216 Wired USB Keyboard Black	
OS	Windows 10 Professional, 64-bit, w/DVD Recovery	
Warranty	3 Year ProSupport with Next Business Day Onsite Service	
Chassis Option	Dell Precision Tower 7810 825W, v2, BW	
Resource Disk	Windows 10 64-Bit OS Recovery and Resource DVDs	
Dell Subtotal		**$5,880.72**
Additional items purchased separately from Newegg.com		
Monitors	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
Mouse	Logitech MX Master Mouse	$69.99
Newegg Subtotal		**$929.97**
System Total		**$6,810.69**

Notes

1. Dell's online configurator does not allow for a 512GB M.2 boot drive, so this system is being specified with a traditional 512GB SATA SSD.
2. The choice of the Xeon i7 E5-1660 v4 is rather disheartening. The 7810 allows up to two physical CPUs, so my original choice would be to get an E5-26xx CPU to be able to upgrade to a second one at a later time. Thus, the choice should logically be the i7 E5-2667 v4, which is almost identical to the E5-1660 v4: A 8-core, 3.2GHz, 20MB cache CPU. The problem is the E5-2667 is a staggering $963.79 upcharge just for that single CPU, for absolutely zero performance benefit.
3. Dell's online configurator allows for up to two graphics cards. We have opted for a pair of M4000s.

The Viz Wiz 8-Core Broadwell E: Newegg Edition

The Viz Wiz 8-Core Broadwell E: Newegg Edition		
Component	**Item**	**Price**
Case	Corsair Obsidian 750D Black Aluminum / Steel ATX Full Tower Computer Case	$149.99
Processor	Intel Core i7-6900K Broadwell E 8-Core 3.2 GHz	1,099.99
Processor Cooler	Corsair Hydro H100i GTX Extreme Performance Water/Liquid CPU Cooler	$104.99
Motherboard	ASRock X99 OC Formula/3.1 Extended ATX Intel Motherboard	$299.99
Memory	CORSAIR Dominator Platinum 64GB (4 x 16B) 288-Pin DDR4 SDRAM DDR4 3333 (PC4 26600) Memory Model CMD64GX4M4B3333C16	$499.99
Graphics	EVGA GeForce GTX 1070 w/8GB	$399.99
	EVGA GeForce GTX 1070 w/8GB	$399.99
	EVGA GeForce GTX 1070 w/8GB	$399.99
Storage	Samsung 960 Pro M.2 512GB Solid State Drive	$329.99
Power supply	CORSAIR AXi series AX1200i 1200W ATX12V 80 PLUS PLATINUM Certified Full Modular Active PFC Power Supply	$309.99
Mouse	Logitech MX Master Mouse	$69.99
Keyboard	Logitech G610 Orion Red Mechanical Gaming Keyboard	$99.99
OS	Windows 10 Professional, 64-bit, OEM	$139.99
Monitor	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
	Dell UltraSharp U2717D 27-inch Widescreen Flat Panel	$429.99
System Total		**$5,164.85**

Notes

1. We are going with the 8-core, 3.2GHz i7-6900K Broadwell E for our Viz Wiz tasks.
2. Because we can use install up to 5 graphics cards in this system, I specified the Corsair 750D full ATX tower case. It has plenty of room for an Extended ATX motherboard and three video cards.
3. I chose the ASRock motherboard for its 5 PCIe x16 slot expandability for hosting multiple graphics cards as well as the inclusion of USB 3.1 ports.

The Viz Wiz Build Analysis

This system has two requirements: Burn through any multithreaded workload across the board in all AEC applications, and render Iray scenes as fast as possible. To that end I went with the Broadwell E 8-core i7-6900K running at 3.2 GHz. It should be overclockable to between 4 - 4.5 GHz with a water cooler.

In looking at what Dell offered in the high end, their Precision 7810/7910 tower models had too many limitations to be competitive. You can choose from almost every E5 CPU that Intel makes, but the upcharge to go to an E5-26xx version, to allow 2 physical CPUs to be installed, was simply crazy high.

Dell was too limiting in other areas as well. It does not offer more than two graphics cards, whereas I was able to get (3) GTX 1070s in the BIY build. The 7810 configuration has the same problem that the 5810 does - you cannot configure a system online with a 2.5″ M.2 boot SSD, and have to settle for a SATA drive.

Interestingly, Dell's minimum RAM configuration is 64GB as 4x16GB modules. Who knows how much more that is internally than Newegg's Corsair 64GB kit.

While some of this may be alleviated by talking with your Dell sales representative, in the end I do not see the 7810 being worth the effort. If you are going to spend over $6,000 for a system, you should be able to get exactly what you need.

XIII. Links

Industry Pressures

Setting up an Amazon EC3 render farm with Backburner
http://area.autodesk.com/blogs/cory/setting-up-an-amazon-ec2-render-farm-with-backburner

Iray benchmarks
http://www.migenius.com/products/Nvidia-Iray/Iray-benchmarks

Revit Specific

Revit Model Performance Technical Note 2017:
http://revit.downloads.autodesk.com/download/2017RVT_RTM/Docs/InProd/Autodesk_Revit_2017_Model_Performance_Technical_Note.pdf

Processors:

Intel's Comparative CPU Database
http://ark.intel.com/

Intel Tick-Tock Model
http://en.wikipedia.org/wiki/Intel_Tick-Tock
http://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html

Intel's Skylake CPU Reviews
http://www.anandtech.com/show/9483/intel-skylake-review-6700k-6600k-ddr4-ddr3-ipc-6th-generation
http://arstechnica.com/gadgets/2015/08/intel-skylake-Core i7-6700k-reviewed/

Intel Skylake (6th Generation) CPUs Compared
http://ark.intel.com/compare/88195,80807,77656,75123

Broadwell E i7-6xxx Series Compared
http://ark.intel.com/compare/94189,94188,94196,94456

Xeon E3-12xx Skylake Series Compared
http://ark.intel.com/compare/88172,88182,88176,88174,88171

Xeon E5-16xx Broadwell Series Compared
http://ark.intel.com/compare/92991,92987,92994,92985,92992

Xeon E5-26xx Broadwell Series Compared
http://ark.intel.com/compare/92983,92989,92979

Intel Core i7 Kaby Lake Mobile CPUs Compared
http://ark.intel.com/compare/95451,95441

Intel Core i7-6xxx Skylake Mobile CPUs Compared
http://ark.intel.com/compare/88967,93341,88969,88970,93340,93336,88972

Intel Xeon E3-15xx Mobile CPUs Compared
http://ark.intel.com/compare/89608,89610,93358,93359,93354

Transactional Synchronization (TSX Instructions) in Haswell:
http://software.intel.com/en-us/blogs/2012/02/07/transactional-synchronization-in-haswell

Passmark CPU Mark High-End CPU Benchmarks
https://www.cpubenchmark.net/high_end_cpus.html

Graphics:

Autodesk Certified Graphics Hardware
http://usa.autodesk.com/adsk/servlet/syscert?siteID=123112&id=18844534&results=1&stype=graphic&product_group=2&release=2016&os=524288&manuf=1&opt=2

GPU Database
http://www.techpowerup.com/gpudb/

Nvidia GTX 1080 Pascal white paper:
http://international.download.nvidia.com/geforce-com/international/pdfs/GeForce_GTX_1080_Whitepaper_FINAL.pdf

Nvidia Tesla P100 Pascal Whitepaper
https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf

BOXX Blogs
http://blog.boxx.com/

BOXX Blogs: GeForce GTX Rendering Benchmarks and Comparisons
http://blog.boxxtech.com/2014/11/17/geforce-gtx-rendering-benchmarks-and-comparisons/

GTC 2016 on demand
http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php?searchByKeyword=&searchItems=&sessionTopic=&sessionEvent=2&sessionYear=2016&sessionFormat=&submit=&select=

GTC 2016 – Advanced Rendering Solution from NVIDIA
http://on-demand.gputechconf.com/gtc/2016/video/s6571-phillip-miller-advanced-rendering-products-for-end-users.mp4

Nvidia mobile GPU comparison sheet:
http://www.Nvidia.com/object/quadro-for-mobile-workstations.html

Memory

Quad-channel RAM vs. Dual-channel RAM
http://www.pcworld.com/article/2982965/components/quad-channel-ram-vs-dual-channel-ram-the-shocking-truth-about-their-performance.html?page=3

DDR4 2133 to 3200 Memory Scaling
http://www.anandtech.com/show/8959/ddr4-haswell-e-scaling-review-2133-to-3200-with-gskill-corsair-adata-and-crucial

Peripherals:

Cherry MX Switches:
http://www.keyboardco.com/blog/index.php/2012/12/an-introduction-to-cherry-mx-mechanical-switches/

Thunderbolt:

https://thunderbolttechnology.net/blog/thunderbolt-3-usb-c-does-it-all

Windows 10:

Download link for in-place upgrade and media creation tool
https://www.microsoft.com/en-us/software-download/windows10