PERFORMANCE BENCHMARKS FOR CONSOLES:

Jonathan Koomey, Kieren Mayers, Joshua Aslan, and James Hendy

Author for correspondence: jgkoomey@stanford.edu, http://www.koomey.com

V33, July 4, 2017

Games consoles are popular devices. Approximately 85 million consoles were sold within Europe over the last ten years[1] – enough for approximately two in every five European households [1]. In 2013 alone, they were estimated to have consumed 6 TWh of electricity in Europe [2], equivalent to the electricity consumption of two million UK homes [3]. As a result, the energy efficiency and climate change impact of games consoles have become concerns for policy makers on an international basis.

In April 2015, the European Commission recognized a Voluntary Agreement (VA) together with console manufacturers to improve the energy efficiency of games consoles under the Ecodesign Directive.[2] Under this VA, manufacturers are committed to ensure games consoles meet targets for maximum power consumption in certain operational modes and minimum automatic power down limits, together with requirements for material efficiency and information reporting.[3] These targets are expected to achieve energy savings of one terawatt-hour per year by 2020 in the EU [4].

Currently, power consumption targets agreed within the VA apply only to media and navigation modes. Measuring the power consumption of such modes is straightforward, as the modes themselves are well defined, meaning test results can be accurately compared among consoles with similar capabilities, with few exceptions. There are many complexities, on the other hand, when attempting to benchmark console performance in active game play.

In 2017, the VA will undergo review, to update the agreement and set new targets for the future. In preparation for this review, console manufacturers must consider "the feasibility of including computational performance in console efficiency benchmarks, where applicable and comparable across devices performing gaming" [4]. If feasible, policy makers anticipate that the development of a gaming efficiency benchmark would allow targets to be set to improve active gaming power consumption, like those established for other modes, and for reporting performance versus efficiency to consumers.

Identifying a suitable metric is a complex task, as the definition of active gameplay is unclear and multifaceted. A wide range of activities fall under active gameplay, and depending on the game, software design, frame rate, video resolution, and system architecture, the power use can

---

[1] http://www.vgchartz.com/analysis/platform_totals/

[2] http://ec.europa.eu/growth/tools-databases/newsroom/cf/itemdetail.cfm?item_id=8239

[3] http://efficientgaming.eu/

vary tremendously. Many games perform computations in the background even if the user is not active, so even the concept of "active game play" may not be clearly defined. Many console games dynamically modify resolution, frame rate, and other image characteristics to optimize the gaming experience for each console platform, depending on the underlying hardware and the gaming software, making gaming performance even more complex and harder to compare between platforms. In addition, user preferences and game design, which are not under the control of console manufacturers, can have a large effect on power consumption in active game play.

The development of computational efficiency benchmarks is not only important for games consoles, but for other products, such as Gaming PCs, where energy efficiency is a topic of concern. For example, Mills and Mills [5] state that "gaming is the most energy intensive use of personal computers" and have conducted pioneering research investigating potentially suitable metrics for PCs, discussed further below. The authors found that the typical enthusiast gaming PC consumes ~1400 kWh/year compared to ~160 kWh/year for the average console, and the aggregate global energy use to be two-times higher for gaming PCs than for consoles. Moreover, they project this gap in demand to widen substantially by the year 2020.

The purpose of this article is to investigate the potential for developing a benchmark to measure the energy efficiency of active gaming across games consoles, in response to the requirement in the console voluntary agreement for the EU.

CREATING CONSISTENT COMPARISONS

Game consoles vary by system architecture and capabilities, and these capabilities change over time. Current generation consoles (like PS4®, PS4®Pro, Xbox One, WiiU, Nintendo Switch, and the forthcoming Microsoft Xbox One X console) have much more powerful graphics and computational capabilities than older generation consoles. Graphics resolution is higher, frame rates are faster, and the overall gaming experience is quite different for these newer machines. In addition, game consoles are increasingly being used to stream video, listen to music, and perform other non-gaming functions. The *computing services* delivered by these devices are simply not comparable to those from earlier consoles.

Even within current generation consoles there are differences in delivered computing services. Game consoles modify frame rates and video resolution depending on the hardware capabilities of each console (to give the best possible gaming experience on each machine). This dynamic nature of consoles makes it difficult to create a truly consistent comparison of computing services (i.e. gaming performance). In fact, there are many dimensions of gaming performance beyond frame rate and resolution. **Table 1** defines some of those factors.

Another interesting subtlety is that current generation consoles, because of their system-on-a-chip design (and other innovations, see [6]) are more "energy proportional" [7] than earlier consoles, and so save more energy when the device is not being used or operating with lower computational output. This makes measurements of efficiency more complicated (because performance and efficiency are both dynamic and varying rapidly over time).

Table 1: Factors affecting gaming performance and user experience

Term	*Definition*	*Note*
Frame rate	Frame rate, also known as frame frequency, is the frequency (rate) at which an imaging device displays consecutive images called frames. The term applies equally to film and video cameras, computer graphics, and motion capture systems. Frame rate is usually expressed in frames per second (FPS). Tearing, stutter, dropped frames, and partially rendered frames can sometimes be an issue, adding more complexity, but at higher FPS rates these issues disappear.	1
Resolution	The display resolution or display modes of a digital television, computer monitor or display device is the number of distinct pixels in each dimension that can be displayed. It is usually quoted as width × height, with the units in pixels: for example, "1024 × 768" means width is 1024 pixels and height is 768 pixels.	2
Anti-aliasing	In digital signal processing, spatial anti-aliasing is the technique of minimizing the distortion artifacts (like rough edges) when representing a high-resolution image at a lower resolution. Anti-aliasing is used in digital photography, computer graphics, digital audio, and many other applications.	3
Tone mapping	Tone mapping is a technique used in image processing and computer graphics to map one set of colors to another to approximate the appearance of high-dynamic-range images in a medium that has a more limited dynamic range	4
Rendering	Rendering is the process of generating an image from a 2D or 3D model (or models in what collectively could be called a scene file) by means of computer programs. Also, the results of such a model can be called a rendering.	5
Special effects	Special effects are created for games by visual effects artists with the aid of a visual editor.	6
Procedural texturing	A procedural texture is a computer-generated image created using an algorithm intended to create a realistic surface or volumetric representation of natural elements such as wood, marble, granite, metal, stone, and others, for use in texture mapping.	7
Scene complexity	Scene Complexity controls the in-game representation of how detailed objects are. A higher setting here results in more complex geometry in things like foliage, rocks, as well as making objects remain highly detailed at farther distances from the player. This is due to LOD (level of detail), which is used to swap lower resolution objects in as the player moves farther away from them and higher resolution objects in as the player moves closer to them. Lower settings result in a less detailed world and objects lose their detail at closer distances to the player.	8
Graphical fidelity	Graphical fidelity can be defined as the combination of any amount of the three things that make up beautiful games (or virtual beauty in general): detail, resolution, and frame rate	9
Dynamic reflections	Dynamic reflections and shadowing move relative to the objects in the game.	10
Visual density	The perceived "visual density" of a screen—and thus the amount of anti-aliasing possibly needed to make computer graphics look convincing and smooth—depends on screen pixel density ("ppi") and distance from the user's eyes.	11

Notes:
1) https://en.wikipedia.org/wiki/Frame_rate
2) https://en.wikipedia.org/wiki/Display_resolution
3) https://en.wikipedia.org/wiki/Spatial_anti-aliasing
4) https://en.wikipedia.org/wiki/Tone_mapping
5) https://en.wikipedia.org/wiki/Rendering_(computer_graphics)
6) None
7) https://en.wikipedia.org/wiki/Procedural_texture
8) https://steamcommunity.com/app/322920/discussions/0/604941528469072612/
9) https://www.reddit.com/r/pcmasterrace/comments/51u8zk/psa_the_graphical_fidelity_triangle_a_visualized/
10) None

11) http://phrogz.net/tmp/ScreenDens2In.html

An additional complexity when comparing game consoles to gaming PCs is that the Graphics Processing Units (GPUs) in consoles are custom designed (omitting some compatibility firmware) and so allow console designers lower level and faster access to the GPU's capabilities than is possible on a gaming PC. GPUs are a significant contributor to both electricity use and gaming performance, and architectural differences among them can't be ignored in attempting to create consistent comparisons.

Overall, a console's power consumption in different modes will depend strongly on GPU utilization, performance, and efficiency. GPU characteristics are, however, not the only determinants of console power consumption and cannot be used to provide a predictable or consistent benchmark (Table 2). Console power consumption is impacted by many other factors such as: CPU, memory, and power supply performance; differences in the functions provided by the operating system; the level of optimization of the firmware; and differences in chip architecture, design, and die-size.

Table 2: Console GPU performance vs power consumption

		GPU performance [1]	Reported power consumption per mode [2] (W)				
					Media		Average
Console	**launch year**	**ce**	Navigation	*Streaming*	*DVD*	*Blu-ray*	gaming
Microsoft Xbox One	2013	1.31	61.0	63.0	68.0	69.0	106.0
Sony PlayStation®4 (launch model)	2013	1.84	77.6	81.9	97.4	89.1	115.1
Microsoft Xbox One S	2016	1.40	27.0	32.0	33.0	33.0	62.0
Sony PlayStation®4 Slim	2016	1.84	44.0	48.4	43.8	48.5	78.9
Sony PlayStation®4 Pro	2016	4.20	60.4	59.3	54.1	59.5	126.1

1. See http://www.eurogamer.net/articles/digitalfoundry-2016-what-the-hell-is-a-teraflop-anyway & https://www.playstation.com/en-gb/explore/ps4/tech-specs/

2. See http://efficientgaming.eu/compliance-reports/product-compliance-report/. Tests for average gaming taken for three top selling games over 5-minute periods.

MEASURING PERFORMANCE AND ENERGY EFFICIENCY

Assessing the energy efficiency of computing devices performing a computing task (like consoles or personal computers) is a challenge. To measure efficiency, we combine a measure of the output of the device (like computations, game play, or a set of consistently defined tasks) with a measure of the electricity needed to deliver that output (typically measured in kilowatt-hours or kWh). This relationship can be characterized using **Equation 1** [8]:

$$Computing\ efficiency = \frac{Computational\ output}{Electricity\ consumption\ to\ deliver\ output} \quad (1)$$

Equation 1 is simple, but applying it to computing devices isn't. Computational output depends a great deal on the computing task, software, and hardware.

For general-purpose computers, performance benchmarks have always engendered controversy. On the one hand, computer scientists rightly worry that performance is strongly influenced by the characteristics of each workload, and it's difficult to define precisely what a generally applicable set of workloads might be for any set of users. On the other hand, high-level comparisons require some benchmark to be used, even if imperfect, and in practice, differences between benchmarks are less important when examining long term big-picture trends, as for example in [9, 10, 11].

Many researchers have wrestled with this problem in the past, including Knight [12, 13, 14], Moravec [15], McCallum [16], and Nordhaus [17]. The work of SPEC <http://www.spec.org> grew out of those early efforts, and it remains a widely-used set of benchmarks that have the imprimatur of industry acceptance. SPEC has many different benchmarks for different applications, and each part of the Information Technology (IT) industry gravitates towards the metrics that are most applicable (or most advantageous) for their application. There are metrics that focus on database queries, metrics that focus on application performance, and metrics that focus on computational speed for CPU based or scientific workloads.

The SPEC workloads were eventually paired with power measurements, at least for servers (https://www.spec.org/power_ssj2008/), growing out of some earlier work [18]. Those measurements (and lots of industry meetings) resulted in what is known as SPEC power, a metric that tied performance measurements for a CPU intensive workload with power measurements at different levels of equipment utilization, resulting in curves that look like those shown in **Figure 1**.

The most important parameters for servers are the idle power (i.e., power use measured with zero computing load) and the maximum power use (measured at maximum computing output). The load curve is typically a straight line between these two points for a server, though of course some computing devices may have workload/power curves with a different shape. Power use and performance are measured simultaneously, so as the computing benchmark is run, power use is tracked, and as the workload becomes more computationally intense, power use generally increases.

Curves of this type characterize the relationship between computing performance and power use. Curves that have high part-load savings (i.e. draw little power at idle) are said to be "energy proportional" [7]. Because most computing activities are concentrated into a small number of hours per year, an energy-proportional computing device will also be an energy-efficient device.

The SPEC power metric has persisted over time (starting in 2007), but is limited to the CPU-intensive SPEC_jbb benchmark. Some in the industry expected SPEC to extend power measurements to other benchmarks, but that has not occurred, and the SPEC power database, while it is still updated by manufacturers, represents the best-in-class servers that manufacturers *want* to benchmark, so it is not representative of typical practice. Nobody forces manufacturers to run SPEC power, so it is widely believed that they just run the servers they expect to do well in the test.

This lack of applicability to the broader market led the EPA's Energy Star server program[4] to commission a new benchmark from SPEC, called the Server Efficiency Rating Tool (SERT). Manufacturers use this tool, found at https://www.spec.org/sert/, to qualify their servers for the Energy Star Servers program. SERT reports similar information to SPEC power, but using a more general benchmark suite of computing activities. There are no current requirements by Energy Star on active computing efficiency for servers, but the program does require the workload/power curve to be created and reported for each server that qualifies for the Energy Star label.

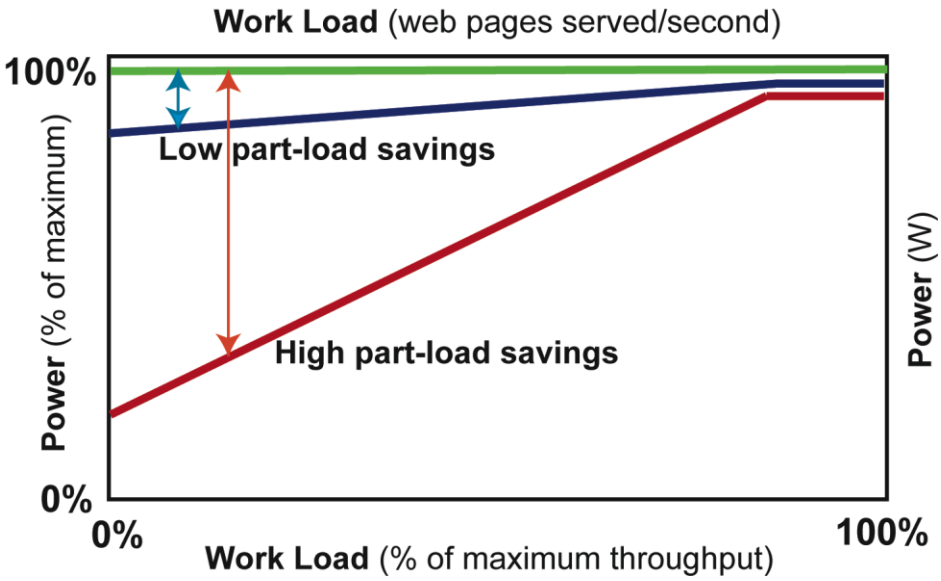


Figure 1: Conceptual Diagram of Energy vs. Computation Metric

Source: Nordman [19].

DEVELOPING EFFICENCY BENCHMARKS FOR GAMING PCS AND CONSOLES

Benchmarking active power efficiency of game consoles is more complicated than for servers. First, the system architectures can vary greatly among console manufacturers, and even more widely when gaming PCs are considered. Second, the concept of "active use", which is clear for a server, may be impossible to define for a console (much console computing happens in the background even if there is no user input or network traffic, and the gaming experience varies significantly across consoles even when considering the same game). Finally, the way games are programmed can have a big effect on power use, with the same game showing widely different power use on different consoles, depending on how much the code is optimized for each platform, the type of game (e.g., sports games vs first-person shooter games) and how frame rates, resolution, and other gaming performance factors are dynamically modified during the game. Because of these complexities, it is unlikely that a curve like Figure 1 can be created for consoles—workload just isn't as uniform (or simple) as it is for servers.

[4] https://www.energystar.gov/products/spec/enterprise_servers_specification_version_2_0_pd

In the preparatory discussions leading up to the voluntary agreements for consoles (2013-2014) there was some discussion of how one might benchmark active compute output, with most attention being paid to measurements of active power when running popular games. The VA currently includes a requirement for signatories to measure this metric and report publicly. In such a scheme, a set of widely used games would be chosen using an objective metric and then power use measured as each game is played, with a focus on just the first five minutes of the game.

Such an approach would be difficult to implement, in part because it would be dependent on characteristics of each game. For example, while some activities in the game may be computationally intensive, other activities may be less so, and power use will vary significantly while playing. The results would vary over time, creating problems for enforcement, because manufacturers would have to retest old models every year using the latest games.

Any protocol for measuring power use under active game play will have to create procedures to ensure tests are consistently applied, repeatable, and representative of actual gaming use. These procedures would also need to be modified over time to reflect the changing mix of popular games and would need to be carefully designed so that electricity use is measured for delivering comparable levels of service (e.g., resolution and frame rates) so that the comparisons between different consoles and gaming PCs are truly consistent ones.

A look at the characteristics of some popular games confirms the complexity of the benchmarking task for gaming platforms.[5] Consider four of the best-selling games for 2015[6]:

1. Call of Duty: Black Ops III – Runs dynamic resolution to try maintain 60 FPS[7].

2. Fallout 4 – Performance issues on both PS4 and Xbox one (Patch 1.03)[8] and Frame rate issues dropping below 30 FPS[9].

3. Star Wars Battlefront – Differing native resolutions (lower on Xbox One)[10].

4. Grand Theft Auto 5 – Lower detail / object density noted for Xbox One[11].

---

[5] Methods discussion for analyzing frame rates at: http://www.eurogamer.net/articles/digitalfoundry-2015-how-we-measure-console-frame-rate

[6] We omitted Madden NFL 2016 (the NPD number two game by unit sales in 2015) because it's a US football-centric game that isn't as widely played in Europe, hence the Eurogamer web site didn't test it.

[7] http://www.eurogamer.net/articles/digitalfoundry-2015-call-of-duty-black-ops-3-face-off

[8] http://www.eurogamer.net/articles/digitalfoundry-2016-fallout-4-patch-improves-console-graphics-quality

[9] http://www.eurogamer.net/articles/digitalfoundry-2015-fallout-4-face-off

[10] http://www.eurogamer.net/articles/digitalfoundry-2015-star-wars-battlefront-face-off

[11] http://www.eurogamer.net/articles/digitalfoundry-2015-grand-theft-auto-5-pc-face-off

Different consoles run different games differently, which shouldn't be surprising. Games are regularly updated by downloadable patches, and a different patch version of a game can affect performance on a console (or a gaming PC). To correctly estimate efficiency in a consistent way would involve correcting for any differences in the quality of graphics output, but since these differences vary dynamically, the calculational and tracking challenge is not a trivial one.

As a proof of concept, **Figure 2** shows power measurements for four popular games taken by Joshua Aslan of Sony in June 2016 on five examples of Sony's PlayStation®4 (all are Model # CUH12xxA). The measurements are taken every second over a five-minute period. The "whisker plots" show maximum, 75th quartile, median, 25th quartile, and minimum values over the measurement period. Taller boxes imply more variation in the data values than shorter boxes.

Appendix A contains the distributions of power consumption for every console and game combination, as well as the time series of power use over time as each game was played on each console. We compare these results using ANOVA statistical tests in Appendix C, which show that the variability observed in the measured power consumption is statistically significant (at the 95% confidence interval) between the console sample used, the sequence of user actions and choices at each stage of a game over time (or phase of gameplay), and the type of game.

Due to the complexity of almost limitless choices, permutations, and combinations of user actions possible within each game, it's impossible to replicate a test exactly. Median, maximum, and minimum power measurements vary for each game title tested when played on different console samples. This demonstrates the difficulty in replicating gameplay (due to the limitless combinations of user actions possible within each game, as well as unseen background functionality not under direct user control) and the statistical variation in hardware and software of the console sample itself. In addition, the plots below highlight the capability of new generation consoles to dynamically scale power consumption as required. Some games, like Call of Duty, show significant power scaling, while others, such as Battlefield 4 (a competing title to Call of Duty), show much less variation.

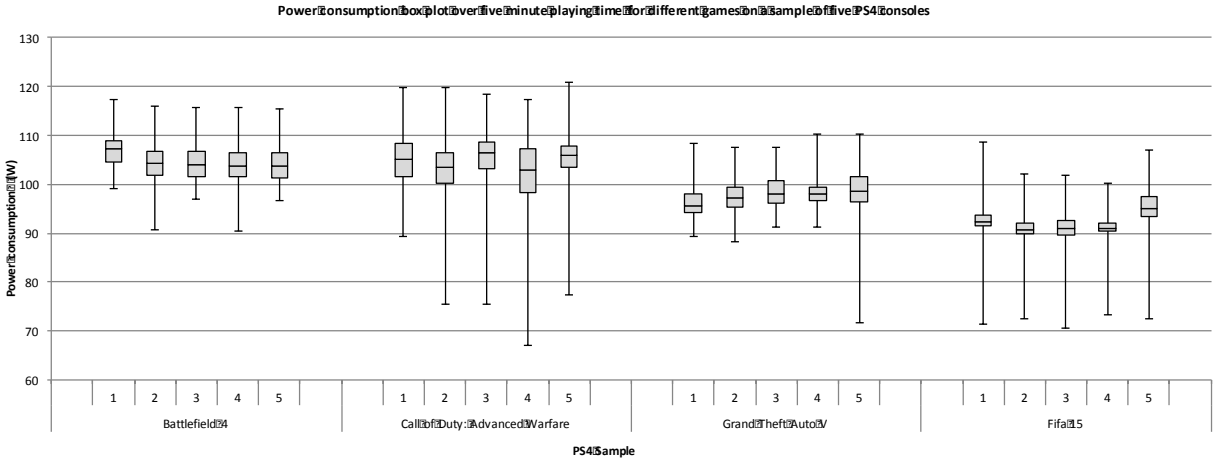


Figure 2: Characteristics of power measurements for four popular games over a five-minute period

Power use even varies significantly when playing the same game on the same console. **Figure 3** shows the same whisker chart as in Figure 2, but with measurements taken when playing one game five different times on the same console (Console 2 from Figure 2). Appendix B shows the detailed distributions and time series measurements for these data, just as in Appendix A. The progress of the game and variations in the way the game story evolves affect power use significantly (verified in Appendix C; Tables C-2 to C-5).

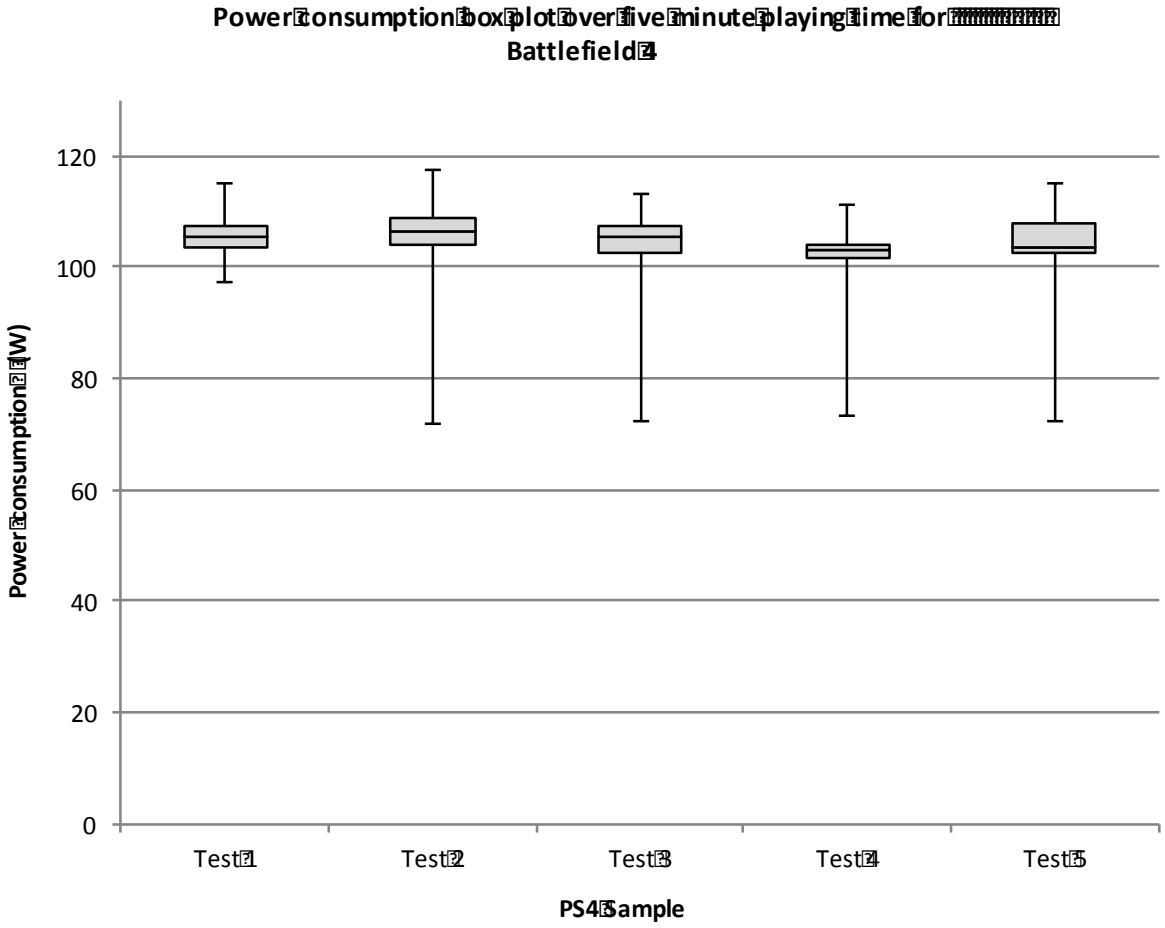


Figure 3: Characteristics of power measurements for one popular game played over a five-minute period five different times on the same console unit

A different approach to benchmarking (distinguished from measuring power levels associated with operating a console) is to give consumers a relative ranking of different products based on component characteristics, which is the approach taken by Enervee.[12] This rating system involves detailed technical knowledge of the hardware specifications in four major subsystems: CPU, GPU, RAM, and hard disk drive.[13] Enervee develops a "performance factor" for each of

[12] https://enervee.com/video-game-consoles/

[13] http://cleantechnica.com/2013/08/02/playstation-4-leads-the-way-in-video-game-console-energy-efficiency/

these subsystems and weights that performance factor equally across the four categories. The Enervee Score takes that performance factor and divides by estimated annual energy consumption, and the result is scaled for all products in the category to cover a 0 to 100 scale.

Enervee's approach gives consumers a credible basis on which to compare the hardware efficiency of consoles, but it is more of a relative informational scale than anything on which a regulation could be based. It also is focused only on hardware, but as shown above, software also has a huge influence on the quality of gaming experience and the level of computational output from a computing system. Ignoring software simplifies the benchmarking task but makes it less likely that a benchmark will be reflective of user experience and actual computing services delivered.

Mills and Mills [5] analyze component-based rated power for gaming PCs, then compare rated power of all components to actual electricity consumption measured while running a GPU frames per second (FPS) benchmark (a benchmark for GPUs of gaming PCs from Unigine: https://unigine.com/products/benchmarks/). They also compare rated to actual component power draws for two CPUs, two GPUs, two motherboards, two power supplies, and three monitors. In addition, they benchmarked the CPUs with Cinebench and examined the effects of overclocking CPUs on performance.

Unfortunately, FPS is not the only measure of graphics performance, never mind gaming performance. In addition, the Unigine benchmark is limited to use with PC GPUs. This benchmark is not technically compatible for use with gaming consoles, because the software layers that allow the CPU to access the GPU in consoles are different than in PCs. On consoles, these layers are less intrusive and more highly optimized, allowing for better performance and energy efficiency for a given GPU and CPU architecture. This also means the system layers needed for a GPU benchmark such as Unigine to run on a PC do not exist on a console, and adding them would result in a benchmark that would not be representative of games console power consumption and efficiency in actual use (because real game play takes advantage of the much faster GPU access the console has, without the interference of the additional system layers in a gaming PC).

A related component-based approach is that used in [20] to create a consistent comparison of energy consumption associated with improving GPUs in gaming PCs. Other examples include the set of allowable total energy consumption adders associated with GPUs of different performance summarized in recently proposed California efficiency standards for computing devices [21] and a 2013 European Union regulation for PCs and servers [22]. Such an approach focuses on an important component – e.g., the GPU- and characterizes a critical parameter affecting performance of that component – e.g., frame buffer bandwidth – or some measurement of performance of that component – like GB/s of data transfers to and from the GPU. Such measures may be relevant for standardized PC architecture, but not for console architectures that are integrated and optimized. Consoles do not have dedicated high bandwidth memory for use with discrete GPUs, but instead use shared high bandwidth memory for use with integrated system components.

WHAT MAKES A GOOD BENCHMARK?

A good efficiency benchmark should be

- repeatable
- representative of real world computing activities
- normalized to equivalent levels of computing services (e.g. frame rates and video resolution, which are related to specifications like HD, Ultra HD, etc)
- comparable in a meaningful and accurate way across platforms (e.g. between types of consoles and between consoles and PC gaming platforms)
- stable over time
- regarded as neutral by competing companies
- based on publicly disclosed test procedures and system settings

The value of a computing benchmark depends on the purpose to which it will be put. Benchmarks have been used for consumer efficiency information, but they have also been used for regulatory proceedings and for utilities to pay incentives to customers to improve the energy efficiency of appliances and electronic equipment. Consumer information represents the least demanding application of computing benchmarks. The bar is higher for benchmarks used in regulatory proceedings or to calculate incentive payments, as it should be. Some efficiency benchmarks are used internally by companies to improve relative efficiency of computing platforms, but are not intended for external consumption.

Below we review the various criteria in the context of existing attempts to benchmark console/gaming PC performance and energy use. These attempts all fall short of what would be needed to create an ideal benchmark, but we can still learn something from each attempt.

Repeatability and representativeness

A reproducible gaming benchmark would require that settings on each device be systematized and recorded. These parameters would include OS/firmware version, game patch version, console system settings (such as native output resolution i.e. 1080P) and in-game graphics settings (if available).

No measurement of gaming performance can be repeated exactly, because game play is dynamic and unpredictable, due to the many possible combinations of actions possible in a game. For this same reason, it's impossible to create a representative computing task for gaming devices in the way industry has done for servers.

Normalized to consistent levels of service

Normalizing to consistent service levels is also impossible, because of the dynamic nature of video resolution, frame rates, and other factors affecting game performance, the complexity of branching choices inside of games, and the multi-faceted nature of the computing services delivered by gaming devices. Industry has attempted to simplify characterization of video services using terms like HD, Ultra HD, or "generations" of consoles within the current version of their VA, but these categories don't reflect differences in all important aspects of gaming

performance. In future, such generational characterizations will need to account for measures of overall console performance beyond image resolution or frame rate.

An additional complexity is that the purpose of gaming is not to produce any specific output (as for servers or computers in business), but *to have fun*. Each person has a unique perspective, and not everything about consoles that can be measured matters to people using the machines. In some cases, changes in console capabilities may not even be visible to users. Given these realities, it is unclear how we can quantify user experience in a consistent and reproducible way.

Comparable across platforms

Because of the differences in the architecture of consoles and PCs, creating a cross platform benchmark has proved to be a challenge. No cross-platform benchmarks that are representative and normalized by level of service currently exist, and it is unlikely that one can be created.

Stable over time

This criterion will never be met exactly, because computing platforms change over time, requiring modifications of benchmarks. But to the extent possible, benchmarks need to remain stable. This criterion shouldn't be hard to meet, assuming industry could agree on a reasonable benchmark. The rate of change in the technology industry makes it imperative to "future proof" any performance metrics to the extent possible.

Vendor neutrality

Even if a test could be designed that is "fair", vendors may object if it disadvantages their product. This implies that a neutral third party would need to design and take charge of the testing.

Based on publicly disclosed procedures

This criterion is relatively easy to meet, and it is in the interest of all stakeholders to release the information so the tests become widely accepted.

CONCLUSIONS

The dynamic nature of consoles creates extreme complexity. It is unlikely that meaningful metrics for comparing gaming performance can ever be developed for game consoles and gaming PCs. The complexity of these devices makes it difficult to define computational output in a way that can be accurately, consistently, and correctly compared across game consoles or between consoles and gaming PCs. Without consistent computational benchmarks, it's unlikely that a benchmark for active gaming will ever be good enough on which to base efficiency regulations or utility incentives to promote more efficient products.

ACKNOWLEDGMENTS

We gratefully acknowledge helpful comments from Leo Rainer, Bruce Nordman, Norm Bourassa, Evan Mills, and Richard Brown at Lawrence Berkeley National Laboratory.

REFERENCES

1. Eurostat. 2015. *People in the EU – statistics on household and family structures - Statistics Explained*. Brussels, Belgium: European Commission. [http://ec.europa.eu/eurostat/statistics-explained/index.php/People_in_the_EU_%E2%80%93_statistics_on_household_and_family_structures]

2. Ricardo-AEA. 2013. *Impact Assessment Study for Sustainable Product Measures: Lot 3 – Sound and Imaging Equipment*. UK: European Commission, DG Enterprise and Industry. Ricardo-AEA/R/ED57346–Issue #1, EC #84/PP/ENT/IMA/11/111131. March 27. [http://ec.europa.eu/DocsRoom/documents/10199/attachments/1/translations/en/renditions/native]

3. IEA. 2016. *Key World Energy Statistics 2016*. Paris, France: International Energy Agency. [http://www.iea.org/publications/freepublications/publication/key-world-energy-statistics.html]

4. Console Manufacturers. 2015. *Energy Efficiency of Games Consoles: Self-regulatory initiative to further improve the energy consumption of games consoles*. Sony Computer Entertainment Inc., Microsoft Corporation and Nintendo Co., Ltd. Version 1.0. April 22. [https://ec.europa.eu/energy/sites/ener/files/documents/Games%20Consoles%20Self-Regulatory%20Initiative%20V1%20-%20Final.pdf]

5. Mills, Nathaniel , and Evan Mills. 2016. "Taming the Energy Use of Gaming Computers." *Energy Efficiency*. vol. 9, no. 2. April. pp. 321–338. [http://link.springer.com/article/10.1007/s12053-015-9371-1]

6. Webb, Amanda, Kieren Mayers, Chris France, and Jonathan Koomey. 2013. "Estimating the Energy Use of High-Definition Games Consoles." *Energy Policy*. vol. 61, October. pp. 1412–1421. [http://www.sciencedirect.com/science/article/pii/S0301421513003923]

7. Barroso, Luiz André, and Urs Hölzle. 2007. "The Case for Energy-Proportional Computing." *IEEE Computer*. vol. 40, no. 12. December. pp. 33-37. [http://www.barroso.org/]

8. Koomey, Jonathan. 2015. "A primer on the energy efficiency of computing." In *Physics of Sustainable Energy III: Using Energy Efficiently and Producing it Renewably (Proceedings from a Conference Held March 8-9, 2014 in Berkeley, CA)*. Edited by R. H. Knapp Jr., B. G. Levi and D. M. Kammen. Melville, NY: American Institute of Physics (AIP Proceedings). pp. 82-89.

9. Koomey, Jonathan G., Stephen Berard, Marla Sanchez, and Henry Wong. 2011. "Implications of Historical Trends in The Electrical Efficiency of Computing." *IEEE Annals of the History of Computing*. vol. 33, no. 3. July-September. pp. 46-54. [http://doi.ieeecomputersociety.org/10.1109/MAHC.2010.28]

10. Koomey, Jonathan, and Samuel Naffziger. 2016. "Energy efficiency of computing: What's next?" In *Electronic Design*. November 28. [http://electronicdesign.com/microprocessors/energy-efficiency-computing-what-s-next]

11. Koomey, Jonathan, and Samuel Naffziger. 2015. "Efficiency's brief reprieve: Moore's Law slowdown hits performance more than energy efficiency." In *IEEE Spectrum*. April. [http://spectrum.ieee.org/computing/hardware/moores-law-might-be-slowing-down-but-not-energy-efficiency]

12. Knight, Kenneth E. 1963. *A Study of Technological Innovation–The Evolution of Digital Computers*. Thesis, Carnegie Institute of Technology.

13. Knight, Kenneth E. 1966. "Changes in Computer Performance." *Datamation*. September. pp. 40-54.

14. Knight, Kenneth E. 1968. "Evolving Computer Performance 1963-67." *Datamation*. January. pp. 31-35.

15. Moravec, Hans. 1998. "When will computer hardware match the human brain?" *Journal of Evolution and Technology*. vol. 1, [http://www.transhumanist.com/volume1/moravec.htm]

16. McCallum, John C. 2002. "Price-Performance of Computer Technology." In *The Computer Engineering Handbook*. Edited by V. G. Oklobdzija. Boca Rotan, FL: CRC Press. pp. 4-1 to 4-18. [http://www.jcmit.com/]

17. Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." *The Journal of Economic History*. vol. 67, no. 1. March. pp. 128-159. [http://nordhaus.econ.yale.edu/recent_stuff.html]

18. Koomey, Jonathan, Christian Belady, Henry Wong, Rob Snevely, Bruce Nordman, Ed Hunter, Klaus-Dieter Lange, Roger Tipley, Greg Darnell, Matthew Accapadi, Peter Rumsey, Brent Kelley, Bill Tschudi, David Moss, Richard Greco, and Kenneth Brill. 2006. *Server Energy Measurement Protocol*. Oakland, CA: Analytics Press. November 3. [http://www.energystar.gov/datacenters]

19. Nordman, Bruce. 2005. *Metrics of IT Equipment — Computing and Energy Performance*. Berkeley, CA: Lawrence Berkeley National Laboratory. Draft LBNL-60330. July 26. [http://hightech.LBL.gov/datacenters.html]

20. AMD. 2016. *AMD Accelerates GPU Energy Efficiency for Gaming PCs*. [http://www.amd.com/Documents/polaris-carbon-footprint-study.pdf]

21. CEC. 2016. *Appliance Efficiency Rulemaking for Computers, Computer Monitors, and Signage Displays*. Sacramento, CA: California Energy Commission. November 23. [http://docketpublic.energy.ca.gov/PublicDocuments/16-AAER-02/TN214560_20161123T144614_15Day_Language_Express_Terms.pdf]

22. European Union. 2013. *Ecodesign requirements for computers and computer servers, Regulation (No 617/2013), Article 2 (20)*. [http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0013:0033:EN:PDF]

APPENDIX A: DETAILED MEASUREMENTS

This appendix shows power use by five different PlayStation® units while playing four different games. **Figure A-1** shows the distribution of power measurements for all combinations of consoles and games, while **Figure A-2** shows the second by second power measurements over time for the same combinations.

Figure A-1: Distribution of power measurements for five consoles playing four games

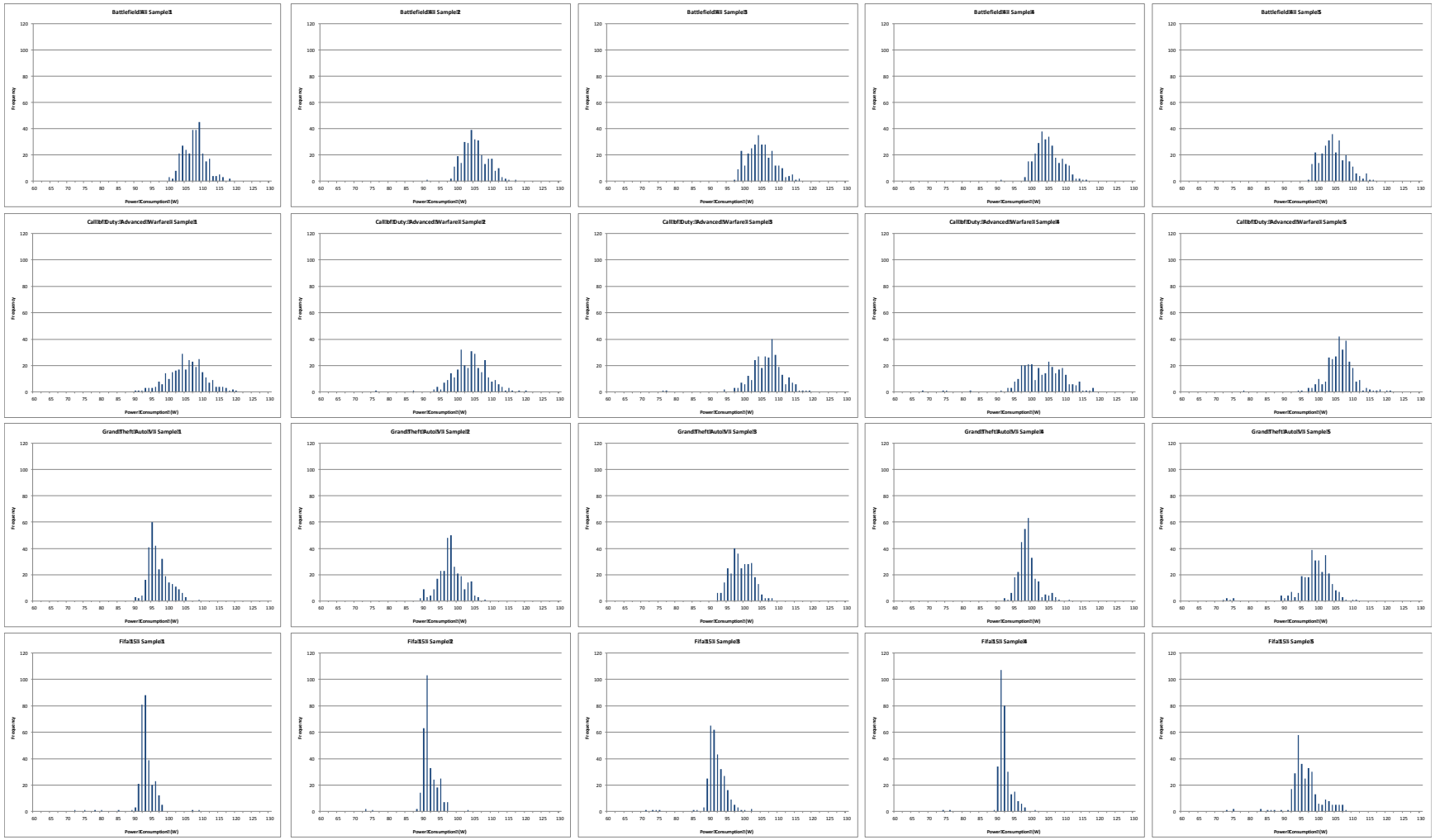
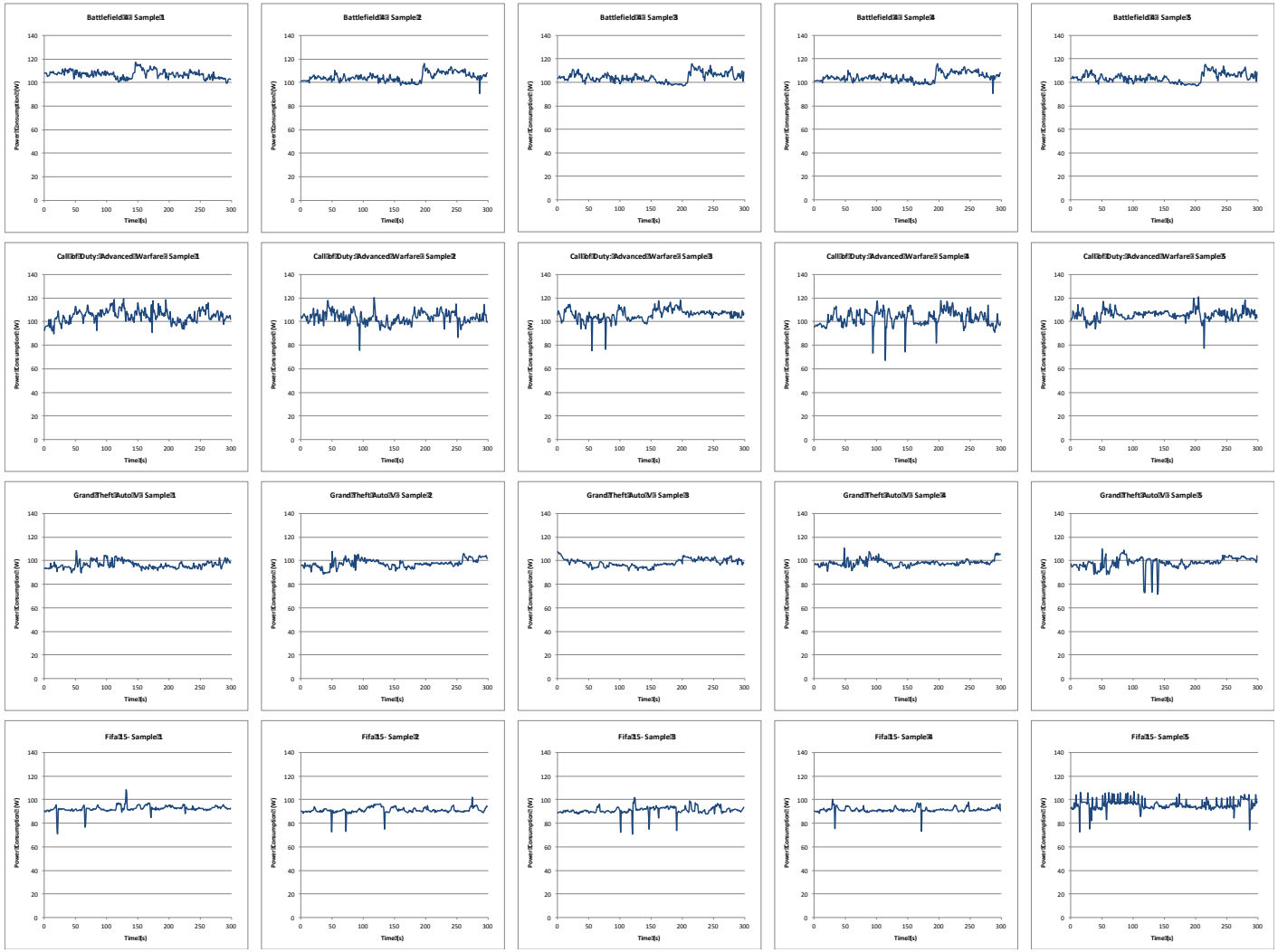


Figure A-2: Time series of power measurements for five consoles playing four games



APPENDIX B: DETAILED MEASUREMENTS OF GAME PLAY ON A SINGLE CONSOLE

This appendix shows power use by the same PlayStation® unit (Console Sample 2 from the figures in Appendix A) while playing the same game (Call of Duty) five different times. **Figure B-1** shows the distribution of power measurements for all five times this console was used to play Call of Duty, while **Figure B-2** shows the second by second power measurements over time for the same combinations.

Figure B-1: Distribution of power measurements for one console playing one game five times

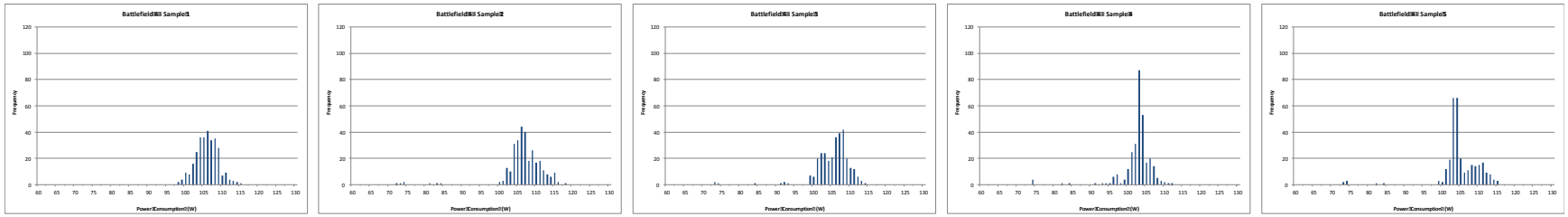
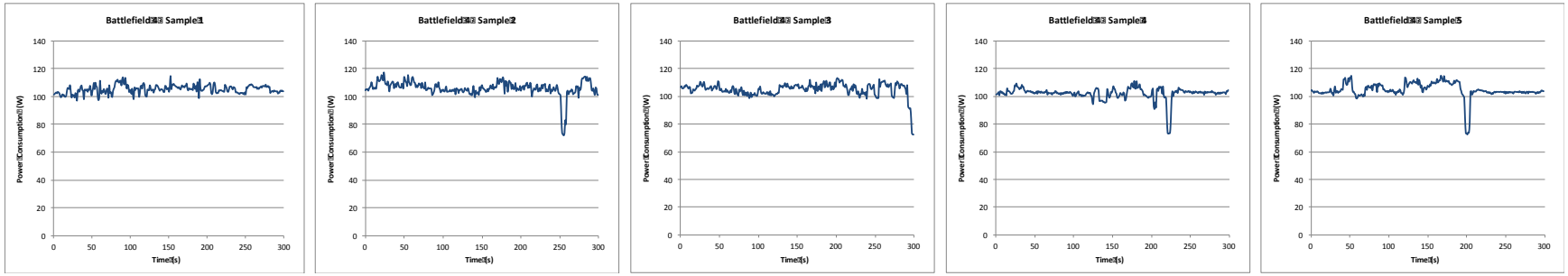


Figure B-2: Time series of power measurements for one console playing one game five times



APPENDIX C: ANALYSIS OF VARIANCE

This appendix details the ANOVA tests for statistical significance between the independent variables of console sample, game used and phase of gameplay on the dependent variable of console power consumption.

All tests are conducted at the 95% confidence interval, $\alpha = 0.05$

1. Console sample and game used

Test used: two-way ANOVA with replication.

Independent variables: console sample and game used

Dependent variable: measured power consumption (sample size of 300, as measurements were made every second for five minutes)

H_0 :

1. there is no significant difference between the measured power consumption of consoles using different samples
2. there is no significant difference between the measured power consumption of consoles using different games
3. there is no interaction between console sample and game used

Table C-1: Two-way ANOVA test for console sample and game used

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	158878.6	3	52959.53	3277.595	0	2.606394
Columns	2741.958	4	685.4894	42.42402	3.85E-35	2.373418
Interaction	7251.822	12	604.3185	37.40046	3.61E-85	1.753788
Within	96625.13	5980	16.15805			
Total	265497.5	5999				

$F > F_{crit}$ and $P < 0.05$ for each case, so we reject all the statements of the null hypothesis.

Interpretation:

Therefore there is statistically significant variability between the console samples tested (using the same game) and between the different games played (on the same console). On top of this, there is a statistically significant interaction between the console sample used and game tested – and power consumption does depend on the type of game tested.

2. Console sample and gameplay phase

To test if the variability due to the period of gameplay – each sample was split into 30 second periods; the first 30s is phase 1, the second 30s is phase 2 etc.

Since we have proved that power consumption has significant variability due to the game used, the impact of time/sequence of action (or "phase" of gameplay) and console sample for each game are tested separately:

Test used: two-way ANOVA with replication:

Independent variables: console sample and gameplay phase.

Dependent variables: measured power consumption

H_0 :

1. there is no significant difference between the measured power consumption of consoles using different samples
2. there is no significant difference between the measured power consumption of consoles during different gameplay phases
3. there is no interaction between console sample and gameplay phase

Battlefield 4:

Table C-2: Two-way ANOVA test for console sample and gameplay phase using Battlefield 4

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	5069.955	9	563.3284	83.31999	1.2E-124	1.886324
Columns	2043.844	4	510.9611	75.57452	3E-58	2.378065
Interaction	5493.683	36	152.6023	22.57089	3.4E-114	1.424915
Within	9803.483	1450	6.761023			
Total	22410.97	1499				

$F > F_{crit}$ and $P < 0.05$ for each case, so we reject all the statements of the null hypothesis.

Call of Duty:

Table C-3: Two-way ANOVA test for console sample and gameplay phase using Call of Duty: Advanced Warfare

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	1958.049	9	217.561	10.35588	1.26E-15	1.886324
Columns	2465.296	4	616.3239	29.33694	1.74E-23	2.378065
Interaction	8733.473	36	242.5965	11.54756	5.58E-57	1.424915
Within	30462.27	1450	21.00846			
Total	43619.09	1499				

$F > F_{crit}$ and $P < 0.05$ for each case, so we reject all the statements of the null hypothesis.

Grand Theft Auto V:

Table C-4: Two-way ANOVA test for console sample and gameplay phase using Grand Theft Auto V

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	3288.483	9	365.387	41.48103	2.58E-66	1.886324
Columns	926.1715	4	231.5429	26.2862	4.54E-21	2.378065
Interaction	3900.752	36	108.3542	12.30105	3.63E-61	1.424915
Within	12772.37	1450	8.808533			
Total	20887.78	1499				

$F > F_{crit}$ and $P < 0.05$ for each case, so we reject all the statements of the null hypothesis.

FIFA 15:

Table C-5: Two-way ANOVA test for console sample and gameplay phase using FIFA 15

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	585.2578	9	65.02865	7.187958	2.9E-10	1.886324
Columns	4558.468	4	1139.617	125.9679	2.32E-92	2.378065
Interaction	1439.366	36	39.98238	4.419463	2.53E-16	1.424915
Within	13117.99	1450	9.046888			
Total	19701.08	1499				

$F > F_{crit}$ and $P < 0.05$ for each case, so we reject all the statements of the null hypothesis.

Interpretation:

There is, therefore statistically significant variability between the console samples tested (during the same gameplay phase) and between the different gameplay phases (on the same console). On top of this, there is a statistically significant interaction between the console sample used and gameplay phase – and power consumption does depend on the gameplay phase (i.e. power consumption varies through each 30s segment of gameplay).