Document Identification	
Identifier	TR-PCEEI-Metrics
Issue	1.0
Date	2021-01-22

White Paper on the Definition of Efficiency Metrics for Computers

What is efficency and how is it measured? – PCEEI

GTD GmbH
Ravensburger Straße 32a
88677 Markdorf

☎ +49 (0) 7544 96440-0
📠 +49 (0) 7544 96440-29

gtd@gtd-gmbh.de
http://www.gtd-gmbh.de

Contents

List of Figures

List of Tables

1 Introduction

In 2019, the European Commission indicated that it would welcome support developing an on-mode test method for computers. CLASP contacted the Commission and offered to assist on this topic, by developing a first version of a software tool that would quantify both energy consumption and performance of different computer configurations across several operating systems. The Commission then invited CLASP and its Team (GTD and Intertek UK) to present their concept to key stakeholders to determine whether this idea should be taken forward.

During the discussion and evaluation of the first iterations of the software tool to measure a computers efficiency, a discussion came up what metric is most suitable to express the efficiency of a computer and whether this metric correctly reflects the complexity of a personal computer consisting of hardware, firmware configuration, operating system and configuration and some user applications.

In this paper we evaluate two metrics – (performance/watt) and (performance/watt-hour) – and explore their mathematical behaviour and suitability as a test method for an energy label for personal computers. In section 2 we explain basic terms such as work, performance, energy (watt-hour), and power (watt) and the best practices to use. In section 3 we present the two metrics, (performance/watt-hour) proposed by Digital Europe and (performance/watt) proposed by CLASP and GTD. We look at properties of both metrics and compare them to the metrics used for other well-known industry efficiency benchmarks. In section 4 we explore different examples, both constructed and from the real world, and discuss how the metrics would behave for both different computers and different operating system configurations on the same computer. In section 5 we will draw conclusions from the theoretical and practical behaviour of the metrics and make a recommendation as to the best metric to use for energy labelling of computers.

2 Basic terms and best practices

This section defines the basic physical terms used throughout this paper when discussing the efficiency metrics and calculation. These definitions are important to ensure that everyone has the same understanding of these basic principles.

2.1 Work

Work, or more specifically for computers, *electrical work* is the work done on a charged particle by an electric field[1]. Moving charged particles by an electrical field may charge up a capacitor (i.e. saving a value to a memory cell in a computer's memory) or switching a transistor (i.e. conducting logic operations in a computer's processor). This means, that work is the result a computer has produced, represented by the contents of the memory and storage, either on the computer itself or on external systems, after a certain number of machine instructions have been executed. This might be, for example, a created document, a transferred amount of data or a rendered frame in a video game. Work does not include any measurement of time.

2.2 Performance

The performance of a computer is the rate at which it can do work. In other words, the amount of work it can do divided by the time it takes to do that work.

$$Performance = \frac{Work}{Time} \tag{2.1}$$

A computer which can do more work in the same time has a higher performance than a computer which can do less work in the same time frame. Or the other way round: A computer which can do the same work in less time, also has a higher performance value compared to a computer which needs more time for a certain amount of work.

Examples of units of performance include: megabytes per second ($\frac{MB}{s}$), frames per second ($\frac{frames}{s}$) or operations per second ($\frac{operations}{s}$).

2.3 Energy

Energy, or more specifically, *electrical energy*, is the energy necessary to produce the electric fields which are necessary to move charged particles in an electrical system. A machine requires a certain amount of energy to do a certain amount of work. Note that this, similar to *work*, does not include any time measurements.

The unit of energy is Joule, or short J.

[1] https://en.wikipedia.org/wiki/Work_(electrical)

2.4 Power

Power is the rate at which energy is required by the electrical device to function according to its specification. In other words, power is the amount of energy needed divided by the time the device is operating.

$$Power = \frac{Energy}{Time} \tag{2.2}$$

The unit of power is thus $\frac{J}{s}$, which is more commonly known as Watt (W).

2.5 Relation to other Fields of Physics

The definition of the four concepts *Work*, *Performance*, *Energy*, and *Power* given above is what is commonly accepted in physics and are concepts that are universally applicable to all sorts of processes in nature and engineering, not just electrical devices or computers.

2.6 Best practices

When working with physics formulas, some best practices should be followed to prevent confusion and detect obvious errors in formulas and results. One such best practice which is particularly important in the context of this paper and the discussion about efficiency metrics is to always include the corresponding units with any scalar number. Keeping track of the units in physics, we can get a hint whether our formula and calculation were correct.

Let's consider a simple example: We want to calculate how far a car, driving at $80\frac{km}{hour}$, can travel in 3 hours. Let's assume we forgot how to solve this problem. What can we do? When we take a look at the units of our numbers it should be immediately clear that the only viable solution is to multiply the two numbers because the result only makes sense if it is a distance, d, commonly expressed in meters or kilometers:

$$d = 80\frac{km}{h} \cdot 3h = 80km \cdot 3 = 240km \tag{2.3}$$

The second step in this calculation simplifies the units so we can see, even before calculating the numerical result that the resulting unit will be km, which is an expected unit for the answer to our problem. Other calculations with the given units and numbers which lead to a different result unit can not be used to answer our initial question of how far the car can travel in 3 hours.

3 Presentation and Theoretical Analysis of the Metrics

3.1 Performance per Power (GTD Metric)

The first metric proposed by CLASP and GTD defines efficiency as

$$Efficiency = \frac{Performance}{Power} \tag{3.1}$$

If we substitute performance and power with what we learned above, and then simplify the resulting expression we get:

$$Efficiency = \frac{Performance}{Power} = \frac{\frac{Work}{Time}}{\frac{Energy}{Time}} = \frac{Work}{Energy} \tag{3.2}$$

The result unit of this efficiency metric is thus the unit of the work (i.e., MB, frames, operations) divided by the unit of Energy, J, so for example $\frac{operations}{J}$.

This metric can be formulated with simple words as the amount of work which can be done with a certain amount of energy.

3.2 Performance per Energy (Digital Europe Metric)

The second metric to be discussed is proposed by Digital Europe and defines efficiency as:

$$Efficiency = \frac{Performance}{Energy} \tag{3.3}$$

If we substitute performance with what we learned above we get:

$$Efficiency = \frac{Performance}{Energy} = \frac{\frac{Work}{Time}}{Energy} = \frac{Work}{Energy \cdot Time} \tag{3.4}$$

The result unit of this metric is $\frac{operations}{J \cdot s}$. In common words this could be described as the amount of work which can be done with a certain amount of energy within a certain amount of time.

3.3 Properties of the two metrics

3.3.1 What does the Metric tell about the Computer?

The most important question to ask is “What does a user gain from knowing the efficiency number?” or “What does the efficiency number tell us about the computer under test?”. One of the main concerns raised by Digital Europe was that the user does not see performance from the GTD metric, in the sense of how long does it take the computer to finish a particular task.

The GTD metric can be explained in the following way: If energy is supplied at a certain rate, the computer does accomplish work at a certain rate. Or alternatively, if a certain absolute amount of energy is supplied, the computer can do a certain absolute amount of work, which is information about the performance of the system under test. So if the user knows the rate at which the computer will do work while meeting the requested power demand, calculating the time needed for a task is easy:

$$Time\ needed = \frac{Work}{Performance} \tag{3.5}$$

It is clear that the efficiency value (Efficiency = Performance / Power) is the quotient of two numbers and does not reveal the individual values of the numerator or denominator, so performance cannot be seen directly. But that is of course also not the intention of the efficiency metric, thus either power or performance has to be given to the user.

In case of energy labels usually all three values are given on the label: efficiency, power and performance. An example is the energy label for fridge/freezer combinations, as shown below. In the label we can see that this device got an efficiency rating of A++ and has enough performance to freeze $54l$ and refrigerate $155l$ at a noise level of $38dB$. Also the expected power demand is $280kWh/year$. Similarly an energy label for a computer could list the performance in frames per second in a standardized video game or the time it takes to render an image of a complex 3D model.

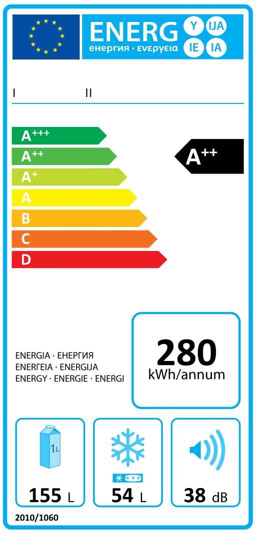

Figure 3.1: Energy label showing efficiency, performance and power demand of a fridge/freezer combination

Now, lets look at the Digital Europe metric (Efficiency = Performance / Energy) using the same approach. According to this metric, a user could expect to supply a certain amount of energy to the computer which makes it run at a certain performance level. Reading this statement, a user might pose one important question which is not answered by Digital Europe's metric: "How long can the computer sustain this performance with the given amount of energy?" or "How long will my battery last before it runs out?".

The efficiency quotient hides the raw values in the same way as with the one calculated by the GTD metric. But, as the efficiency label does not report the time needed to run the test, the user is not able to reverse the calculation which was done to get the efficiency value, and determine the performance and energy consumption.

Without an answer to the question posed above, the efficiency rating doesn't make sense as the user still knows nothing about the computer. The answer to those questions is the efficiency as given by the GTD metric.

3.3.2 Should we all use microcontrollers instead of personal computers?

One concern raised by Digital Europe about the GTD metric is that it will award a higher efficiency value to a microcontroller than to a computer. A microcontroller is a computer on a single silicon chip, containing processor cores, memory, flash storage, and other integrated circuits. Microcontrollers are usually only used for specialized tasks in embedded contexts, and can thus not easily be compared to personal computers in the first place.

Neverthelss, a limited comparison between personal computers and microcontrollers is possible when only looking at tasks (for example adding two numbers) which can be done by both a microcontroller (e.g., as used in an electronic calculator with a seven segment LCD) and a computer. For this particular task, the higher efficiency value would be achieved by the microcontroller, because it uses less energy to add the two numbers together.

That said, the concern will not be be a problem for several reasons:

- No one plans to label microcontrollers and compare them to personal computers
- Microcontrollers do not fulfill the same of a users needs and do not give the expected experience - so even with an A+++ rating, no one would buy a microcontroller to replace a personal computer (they don't offer the same functionality).
- The proposed worklets set a baseline for performance - no microcontroller can run any of the worklets, so its impossible to compare them to personal computers with the method proposed by GTD even if it were desired.
- It is expected that the worklets have to be reevaluated after a certain amount of time to accommodate for changes in technology or when new functionality is becoming relevant in personal computers. This will ensure that the baseline for performance is set adequately.

The lines between microprocessors used in desktop computers, system-on-chip designs as used in smartphones and tablets and microcontrollers are somewhat blurry and new products are released to the market every day and the feature sets overlap in some parts, which leads to technologies emerging today which were previously considered to be too low powered to replace a personal computer, such as the ARM-based Raspberry Pi single board computers, which can fulfill office tasks, some multimedia tasks such as video streaming or even are used as retro-gaming consoles.

Another example worth mentioning are the new ARM devices recently released by Apple[2]. GTD has not tested such a device yet, but from what can be read on the news, they seem to be capable machines, fulfilling a users needs, with exceptional performance and battery life (and

[2] https://www.apple.com/newsroom/2020/11/apple-unleashes-m1/

thus presumably very high energy efficiency). These new computers should not be penalized for optimizing battery life over performance. Although there is still a substantial difference between a typical microcontroller and an Apple M1 processor or a Raspberry Pi, they are somewhat closer to microcontrollers than traditional microprocessors of the x86 architecture. If the traditional manufacturers of microprocessors cannot keep up with these new products, it is not the fault of the efficiency metric.

3.3.3 Independence of the specific worklet length

Worklets for testing computer efficiency define a certain amount of work to be done by the computer. The type and amount of this work is relatively arbitrary and can be discussed – one might argue that rendering 100 frames in a video game is a very small amount of work and that it would be more useful to render 10000 frames instead. Another person might feel that even 10000 frames is way too low as modern graphics card can easily render hundreds of frames per second in some games and to get repeatable numbers a larger amount of frames should be rendered.

Given that there are performance differences to be expected during the course of a run of a worklet, each worklet needs to be engineered in a way to produce reliable results even though caching, system temperature and boost frequencies may change. This means mostly that the worklet must not be too short, as at the beginning of a test the system may reach higher boost clocks but data is not yet cached. Combining these two effects, their impact on performance being opposite to each other, affect performance either positively or negatively depending on the workload.

Once the worklet is long enough to average out the effects of caching and thermal parameters have reached an equilibrium, the efficiency of a computer is not expected to change by a lot as the system has a constant configuration and a constant workload. Or in other words, the efficiency value produced should be the same regardless of whether the system is rendering 1000 or 10000 frames.

It is obvious that for an increased amount of work, both the energy and amount of time increase. This means that for both efficiency metrics, the numerator and the denominator increase. The table below reproduces the two efficiency metrics for ease of comparison.

Table 3.1: GTD and Digital Europe Efficiency Metrics

Metric	Equation
GTD Efficiency Metric	$Efficiency = \frac{Performance}{Power}$
Digital Europe Efficiency Metric	$Efficiency = \frac{Performance}{Energy}$

The GTD efficiency metric still produces the same efficiency value because numerator and denominator both rise proportionally to the increased amount of work. The Digital Europe metric, on the other hand, has two numbers in the denominator that both increase proportionally with the increased workload. This means the denominator is growing quadratically and thus much faster than the numerator which grows linearly, resulting in the efficiency value decreasing as the amount of work done during the worklet run increases.

Taking this to the extreme, with a Worklet which runs forever, we can see that the Digital Europe metric does not produce a useful number anymore – all computers have an arbitrarily small efficiency value:

$$\lim_{Work,Energy,Time \to \infty} \frac{Work}{Energy \cdot Time} = 0 \tag{3.6}$$

The above equation can be interpreted as: "the longer we test a computer system, the lower its efficiency score."

Furthermore, the efficiency value shall be a property of only the system under test and nothing else, as the system under test is awarded an energy label in the end. But as we have seen from our theoretical discussion above, the worklet selection will affect the resulting efficiency value when using the Digital Europe metric, which makes the metric not only a property of the computer under test but also a property of the test software. This could be considered as a serious flaw in the labelling system.

Finally, the Digital Europe metric reduces the maintenance options and sustainability of the test software. As we have discussed above, it may be necessary to change the amount of work a worklet executes in the future to compensate for caching and thermal effects. When using the Digital Europe metric, a recalibrated worklet would not produce results that comparable to the old version. With the GTD metric however, the values would still be comparable.

3.3.4 Rewarding Future Improvements to Performance and Energy Consumption

It is quite obvious that future computers will provide better performance and efficiency. This trend has been driving the computer industry for decades and efficiency of computers has increased and is projected to increase further in the future.

As energy labels are a measure to protect the environment and fight climate change by guiding consumers which device to buy, it is necessary to discuss whether improvements to performance and improvements to power demand/energy consumption should be rewarded equally or if improvements to one of the two is preferable over the improvements to the other. This could also be expressed as: What is the costly resource we have to try conserve? Time or Power/Energy? However, the two efficiency metrics do not behave in the same way regarding this topic.

The GTD metric rewards improvements to both performance and power demand equally. This means that an improvement to performance of factor two doubles the efficiency score in the same way as an improvement to the power consumption by factor one-half. This is easily understood as the performance value in the GTD efficiency metric only depends on performance of the computer and the power value only depends on power demand.

The Digital Europe metric rewards improvements to performance more than improvements to power demand/energy consumption. This is due to the fact that in this metric, the performance value in this efficiency metric only depends on performance, but the energy value depends on **both** performance and power demand. Thus the increased performance is rewarded twice.

Rewarding performance more than lower power demand is not desirable for an energy label meant to reward efficiency. It distorts the metric and biases the results, favouring performance over power demand.

3.3.5 Rewarding Different Configurations of the Computer System Under Test

A single computer system can be configured by the user in various ways for example the Microsoft Windows operating system offers different power profiles for "Maximum Performance", "Battery Saving" and "Balanced". Similarly Linux provides different adjustments for the CPU governor, called "performance", "powersave", and "ondemand".

When a computer system is tested for efficiency in each of these modes, it is expected that differences in power demand/energy consumption and performance will be observed. As with every physical system, be it mechanical or electrical, when demanding peak performance, efficiency is not optimal due to excess heat generation. This becomes even more apparent when overclocking a computer: To get rid of all the excess heat a water cooling solution is needed whereas on a system running at lower clock speeds, it can be enough to have a small air fanor even be passively cooled.

In this way, it can be expected that a mode called "Battery Saving" will limit the clock speed of the CPU, resulting in slightly less performance, but increasing the user-perceived efficiency (i.e., the battery lasts longer).

Another aspect to take a look at is power saving states built into the CPU, called C-states. Modern CPUs have up to 10 C-states which power down parts of the CPU to save energy when these parts are not in use. C-states can be disabled by the user, at least on the Linux operating system, so it would be interesting to test a system with a different number of C-states enabled and see how the two efficiency metrics reflect these changes. It is expected that disabling of C-states will lead to slightly increased performance as the CPU does not spend time to wake up when it is time to do work but obviously this increases power demand / energy consumption during idle periods because the CPU cannot turn off unused components.

As power profiles and C-states affect performance and power demand/energy consumption at the same time to various degrees, the behaviour of the metrics can not as easily be deducted from the formula as we did before. We therefore discuss this point in detail in § 4 with some concrete examples.

3.4 Metrics Used in Other Benchmarks

3.4.1 SPECpower

The SPECpower benchmark describes its efficiency metric in a methodology paper[3], chapter 7:

> Performance/Power Metrics
>
> All power-measured benchmarks will have at least two distinct measurement segments: Benchmark at full performance and Active-Idle after benchmark. Some benchmarks will also include intermediate throughput intervals in the benchmark definition. The average power for each distinct measurement segment must be reported.

[3] https://www.spec.org/power_ssj2008/docs/SPECpower-Methodology.pdf

As we can see the time a worklet needs to run is not even reported by SPECpower, only performance and power.

It also describes the comparability of this "Performance per Power rating":

> Each benchmark that is defined to include power measurements will have a DIFFERENT Performance per Power rating, both because the Performance per Power calculation depends on the throughput values for the benchmark involved and because each benchmark targets a different business model. As such, the metric should be labeled in such a way as to discourage comparison across benchmarks. Terms like "ssjPerformance-per-Power" and "mailPerformance-per-Power" or "webPerformance-perPower" appropriately separate the power metrics from different benchmarks.

SPECpower thus uses a metric which is very comparable to the GTD metric. There are no other alternative metrics discussed in the SPECpower methodology paper.

3.4.2 SERT

SERT describes the calculation of its metric in a separate paper[4], chapter 3:

> The SERT 2 metric, also called SERT 2 Efficiency Score, is a final aggregate of all the power and performance values measured during a SERT run. It is designed to enable regulators to make a decision on whether or not to apply an energy label to a tested system. The SERT suite calculates the SERT 2 metric from the separate workload scores, which in turn are aggregates of all efficiency scores of the worklets within the given workload. The SERT 2 Efficiency Score is a single number that indicates the overall energy efficiency of the tested system.

The SERT paper also gives the explicit formula:

$$Eff_{load} = \frac{Normalized\ Performance}{Power\ consumption} \tag{3.7}$$

The normalization of performance in SERT is done to compare servers with different amount of CPUs or memory channels, corresponding to the type of worklet.

As we can see, SERT also uses an Efficiency Score which is based on power and performance values, and is comparable to the GTD efficiency metric. The execution time of the worklet is not measured. There are no other metrics discussed in the SERT metrics paper.

3.4.3 Phoronix Test Suite

Phoronix Test Suite, the open-source testing framework used by the testing approach developed by GTD also uses the efficiency metric as proposed by GTD. Phoronix Test Suite is not only used by its creators for the benchmarks they do on their website[5] but also by other computer review and testing websites, trade magazines, etc.

[4] https://www.spec.org/sert2/SERT-metric.pdf

[5] https://www.phoronix.com

3.5 Other relevant work in science

We looked to find if there was literature that used a metric similar to the Digital Europe metric. We found a paper by Gonzalez and Horowitz, examining efficiency of CMOS circuits[6].

This paper defines a metric called Energy Delay Product (EDP) which is used for calculating the operating point for CMOS chips. EDP is comparable to the metric proposed by Digital Europe, as it uses not only the energy needed for the CMOS chip to carry out an operation but also the time delay introduced by this operation.

An important difference between these considerations and the ones presented in this paper is that the result of the operation of a CMOS chip is fixed, in contrast to a Worklet where the amount of work is defined. So for CMOS chips we don't run into the problem that if an operation takes a very long amount of time, it could cause the metric to trend towards zero regardless of carried out work and power used.

Other even more exotic metrics are defined in a paper by Mahmud[7], which first of all criticises the EDP metric for over-rewarding performance improvements, similar to what we presented above. Alternative metrics according to Mahmud are Powerup, Speedup, and Greenup. These metrics are then used to classify different software algorithms solving the same problem (e.g. sorting an array of numbers). Since we do not do comparisons between worklets doing the exact same work, but different computer systems running the same workloads, the usefulness of these metrics is also limited.

Apart from the mentions in the Gonzalez and Horowitz paper, there are no users of the EDP metric that we are aware of, nor any of the industry benchmark tools used in production.

[6] https://web.stanford.edu/class/archive/ee/ee371/ee371.1066/handouts/gonzalez_97.pdf

[7] https://greensoft.cs.txstate.edu/index.php/2019/12/12/using-the-greenup

4 Examples

4.1 Independence of the specific worklet length

We want to give some numerical examples for the theoretical discussions presented above. Fortunately, Digital Europe kindly provided several examples.

First, define a baseline computer which delivers a performance of $2000\frac{operations}{s}$, at a power consumption of $10W$ and is able to complete the requested hypothetical task within $2h$:

Table 4.1: Baseline system

	Baseline
Perf (op/s)	2000
Power (W)	10
Time (hr)	2
Energy (Wh)	20
Perf/W	200
Perf/Wh	100

If we now modify the same worklet to do twice as much work, and assume that work will take twice as long and will use twice as much energy, we get the following result:

Table 4.2: Baseline system and increased work

	Baseline	**Increased work**
Perf (op/s)	2000	2000
Power (W)	10	10
Time (hr)	2	4
Energy (Wh)	20	40
Perf/W	200	200
Perf/Wh	100	50

We can see that the GTD metric (Perf/W), has not changed at all when the worklet is modified in this way, but the Digital Europe metric though went from 100 to 50, meaning the efficiency of the system under testing has been cut in half without making any change to the hardware or configuration of the computer under test.

4.2 Rewarding Future Improvements to Performance and Energy Consumption

When we now start changing the performance and power demand values of our hypothetical computer, we can see how the two metrics reward these improvements. For this scenario we will assume that two competing computer manufacturer set different priorities in the development effort. Manufacturer A increases performance, while power consumption stays the same, manufacturer B decreases power consumption while the performance remains constant.

For manufacturer A (performance improves, power same) we might get the following:

Table 4.3: Baseline system and increased performance

Manufacturer A	Baseline	Increased performance
Perf (op/s)	2000	4000
Power (W)	10	10
Time (hr)	2	1
Energy (Wh)	20	10
Perf/W	200	400
Perf/Wh	100	400

For manufacturer B (performance constant, power decreased) we might get the following:

Table 4.4: Baseline system and decreased power demand

Manufacturer B	Baseline	Decreased power demand
Perf (op/s)	2000	2000
Power (W)	10	5
Time (hr)	2	2
Energy (Wh)	20	10
Perf/W	200	400
Perf/Wh	100	200

For both manufacturers the computer efficiency is improved by a factor of 2 (i.e., work doubled for same power, power halved for same work), and we can see the computer of manufacturer A is rated at 400 according to the Digital Europe metric, while the computer of manufacturer B is only rated at 200. However the GTD metric shows that both manufacturers are awarded the same (doubling) efficiency value of 400. We can conclude that the Digital Europe metric penalizes manufacturers who do not focus on performance improvements as the top priority.

A similar comparison can be made by looking at cars. If we look at a low cost sedan and compare it to a sports car, we can see a similar behaviour. We have compiled a table showing estimated data for these two cars.

For travelling 100 km with each of the cars we get the following numbers:

Table 4.5: Car comparison input values

Car	Avg. Consumption (l/100km)	Power (liter/hour)	Time (h)	Performance (km/h)	Total energy (liter)
Sedan	6	6	1.00	100	6
Sports Car	16	44.8	0.36	280	16

If we now calculate the efficiency scores for both cars we get the following results, with the most efficient car marked in bold for each metric:

Table 4.6: Car comparison result values

Car	Performance per Power	Performance per Energy
Sedan	**16.667**	16.667
Sports Car	6.25	**17.5**

We can see that the sports car is rated as more efficient than a sedan according to the Digital Europe metric. However, the sports car is less efficient, because it is optimized for performance and not efficiency. If you can afford a sports car you usually do not care much about efficiency because the fuel cost doesn't matter, but performance can be felt directly while driving so it matters for some people. Most people though would prefer a car which has enough performance to do the required work (i.e., commuting to work every day, shopping for groceries, ...) but also a car that is cheaper in the long run, with a lower total cost of ownership. In this example, that is the sedan.

The same situation exists regarding computers: A gamer who wants to play the latest games at the highest resolution, will always buy the computer which delivers the needed performance. But the average user should be guided towards a computer which has a good balance between performance and power demand and thus efficiency. Not everyone should be incentivized to buy a luxury sports car.

4.3 Rewarding Different Configurations of the Computer System Under Test

4.3.1 Power Profiles

Another approach to comparing the two efficiency metrics is to try different power profile configurations on the same computer, and compare the results of the test runs. To evaluate this analysis, we used a 2018 HP laptop computer, and worked with the two pre-set power modes: long battery life and high performance. The following two tables presents the results of the runs conducted for each of these power modes.

Table 4.7: Test results for MS Windows "Max Performance Mode" profile

Worklet	Performance	Average Power (W)	Energy Consumed (Wh)	GTD Efficency Score	DE Efficiency Score
libjpeg	162.39 MPix/s	27.8	0.165	**5.841**	**982.66**
OSBench	1263.33 us/Event	31.07	2.146	0.255	3.689
AOBench	71.27 s	28.13	1.671	4.987	**83.974**
IOzone	387.88 MB/s	27.2	0.252	**14.260**	**1541.19**
t-test1	220.14 s	28.2	5.173	1.611	8.781
FreeCAD	19.04 s	29.27	0.464	17.946	1131.23
LibreOffice	2.355 s	29.3	1.279	144.92	**3319.72**
Geometric Mean	–	–	–	6.408	**205.483**

Table 4.8: Test results for MS Windows "Battery Saving Mode" profile

Worklet	Performance	Average Power (W)	Energy Consumed (Wh)	GTD Efficency Score	DE Efficiency Score
libjpeg	161.97 MPix/s	27.73	0.166	5.840	977.90
OSBench	1240 us/Event	27.43	1.861	**0.294**	**4.334**
AOBench	80.61 s	24.47	1.644	**5.070**	75.479
IOzone	273.77 MB/s	19.83	0.323	13.804	847.42
t-test1	222.61 s	23.13	4.291	**1.942**	**10.468**
FreeCAD	21.17 s	20.84	0.368	**22.674**	**1285.23**
LibreOffice	2.72 s	22.63	1.126	**162.44**	3265.41
Geometric Mean	–	–	–	**7.043**	197.951

A higher performance value means a better performance for the following worklets: libjpeg and IOzone, whereas a lower performance value means a better performance for the following worklets: OSBench, AOBench, t-test1, FreeCAD and LibreOffice. For the presented geometric mean, a higher value means better efficiency.

As we can see, the Digital Europe metric classifies the "Max Performance" power profile as the most efficient one whereas the GTD metric classifies the "Battery Saver" profile to be more efficient. This outcome is further evidence of the concern raised in this paper, that the Digital Europe metric prioritizes performance over efficiency whereas the GTD metric prioritizes efficiency (i.e., a "Battery Saver Mode" is expected to shift the computer to its most efficient configuration in order to extend battery life). For both metrics, the difference in geometric means between configurations is above 3%, and thus most likely not a measurement error.

For the individual worklets we can see that not for every worklet an increase in efficiency is shown by either metric when switching to battery saving mode, but in case of the GTD metric, numbers are at least very close (for libjpeg the efficiency can be considered identical, and IOzone changes from 14.2 to 13.8 which is only a 3% difference and within the margin of measurement error). For the Digital Europe metric, we see much bigger discrepancies, e.g. IOzone is reported nearly twice as efficient in "Max performance mode".

"Battery Saver" mode certainly decreases performance, which we can clearly see from the results – for the GTD metric it also increases efficiency at the same time. In contrast to that, the Digital Europe metric does not record an overall increased efficiency in Battery Saving mode, instead, it simply appears to slow the user down by delaying the results the user is waiting for. One would assume this was not the intention of the engineers who developed the "Battery Saver" mode at Microsoft.

As shown in the table, the GTD metric shows that efficiency increases in "Battery Saver" mode, which is exactly what a user expects when switching to this mode – e.g., when battery is about to run out. In this situation, the computer does everything possible to be more efficient (sacrificing some performance) to allow the user to continue working as long as possible before the battery is finally depleted.

4.3.2 C-States

Next, we also looked at how a computer behaves when we disable the C-state power saving mechanism of a CPU compared to it's default settings (which includes C-state power saving programme being active). We will take the same approach as done above for the different user-selectable power profiles. When a processor has C-states enabled, it can dynamically disable unused components, governed by the software designed to optimize efficiency based on the task. But if some or all C-states are disabled, it will consume more power due to more of its components being switched on regardless of the tasks – thus there is no longer any optimization and power is wasted. The table below presents a comparison of these two scenarios, with all C-states enabled and only with C-state 1 enabled.

Table 4.9: All C-states enabled

Worklet	Performance (s)	Calculated Time (s)	Average Power (W)	GTD Efficency Score	DE Efficiency Score
OSBench	40.31	5.75	16.74	**14.82**	**9280.64**
AOBench	38.78	38.25	17.34	**14.87**	**1399.87**
T-test1	27.78	27.25	19.48	**18.48**	**2442.04**
FreeCAD	16.36	16.50	19.23	**31.79**	6936.81

Table 4.10: Only C1 enabled and higher level C-states disabled

Worklet	Performance (s)	Calculated Time (s)	Average Power (W)	GTD Efficency Score	DE Efficiency Score
OSBench	39.93	6.75	17.40	14.40	7678.15
AOBench	39.03	39.00	18.55	13.81	1274.84
T-test1	27.15	27.00	20.76	17.75	2366.08
FreeCAD	15.85	15.75	19.94	31.65	**7233.94**

As we can see in this second example, the Digital Europe metric sometimes prefers the configuration with most C-states disabled, as it overcompensates improved performance. This is the case for 1 out of 4 tested worklets.

The results for the GTD metric are more consistent, although the differences observed are very small, as the performance difference between the two test configurations is not as pronounced as it is with the previous example, based on the power profiles. Nevertheless, it is trending in the right direction – the GTD metric rewards software that adjusts the CPU C-states to capture energy savings while the Digital Europe metric does not necessarily capture and reflect C-state power saving mechanisms.

5 Conclusion

Throughout this paper we have discussed the two efficiency metrics in detail from a theoretical point of view as well as practical point of view. In all examples presented, the Digital Europe metric produced values that were not representative or proportional to the changes being made either to the Worklets or to the computer under test. The reason for this distortion is due to computer performance being weighted more heavily in the Digital Europe metric ($\frac{Performance}{Watt \cdot hour}$) while in the GTD metric ($\frac{Performance}{Watt}$) it is given equal weighting with power.

In a worked example, we demonstrated with simple math that when a manufacturer redesigns a computer to keep the same performance but cut power consumption in half, or redesigns a computer to double performance at the same power consumption, the two metrics do not reward this doubling in overall efficiency in the same way. The GTD metric recognizes each of these effects as a doubling in efficiency, however the Digital Europe metric strongly favours the manufacturer to doubled performance and fails to reward the one who maintained performance but halved power consumption.

One of the major concerns Digital Europe raised when evaluating the GTD tool was that it might not take into account the sophisticated power saving mechanisms developed by component manufacturers and system integrators. One example of these power saving mechanisms is a feature that disables CPU components to save energy when there is an idle period or when the processor isn't being fully utilized (C-states). The other example given where different power profiles changing the configuration of the operating system. GTD has demonstrated now that when running the Worklets, which are using real life software and actual tasks that are performed on computers, run the processor at different power levels and thus these power saving mechanisms are represented and captured. The Digital Europe metric, however, does not reward these features because it places an excess of emphasis on performance.

Ultimately, what is needed for establishing an energy efficiency label for computers is an efficiency metric that treats performance improvements and power saving improvements equally. If a manufacturer doubles performance while holding power constant or they maintain performance while halving power, both should result in the same (i.e., equal) improvement on any energy-efficiency labelling scale. In the next phase of this work, extensive testing will be performed which will underscore the appropriateness of what we are calling the 'GTD metric' in this paper.