LEAP Track 1 'Powermanagement' Pilot analysis

Created by Certios/WCoolIT
For LEAP
commissioned by Netherlands Enterprise Agency

AUTHORS

Dirk Harryvan Certios B.V.
Marco Verzijl WcooliT B.V.
Max Amzarakov WcooliT B.V.

Greenit
amsterdam region

PREFACE

"This report is dedicated to Mees Lodder. The ideas about the server idle coefficient that Mees Lodder developed together with Dirk Harryvan formed the basis of the LEAP project. Mees died in December 2019."

EXECUTIVE SUMMARY

Over the past decade, ICT energy efficiency programs have produced enormous advancements in the Netherlands. Because of this attention for energy efficiency, datacenters in the Netherlands have vastly improved their PUE, most datacenters are now highly efficient from a facility standpoint.

At the same time, ICT hardware manufacturers have continued to develop servers, storage and network equipment resulting in enormous advances in the performance per kWh. These advancements have been characterized as Moores law or also as Koomeys law. In addition, Servers in particular have been imbued with features that allow the machines to match energy use to workload.

The combined advancements of facility and ICT efficiency have however not led to a decrease of energy use for ICT services in the Netherlands. It is generally assumed that the ICT sector still exhibits a moderate increase in energy demand, fueled by a tremendous increase in the use of ICT services.

The LEAP ("Lower Energy Acceleration Program") is concerned with improving the energy efficiency of ICT services. LEAP is run by a core team and with a coalition of parties across the datacenter value chain.

The first track of LEAP was sparked by the observation that the overall electric power draw of datacenters is almost constant. This is in sharp contrast to the know variations in demand for ICT services. As a result, an investigation has been started to try to find the reason for this stable power draw as well as to try to lower the overall energy use of datacenters by introducing a more workload connected power draw. The LEAP core team mobilized a group of companies willing to take part in pilots to provide data on their current ICT environment and willing to change certain settings to measure the effect on energy use of the servers being monitored.

The following document describes the results obtained in the LEAP Track 1 'Power management' Pilot analysis, this final report contains additional measurements collected to increase the overall reliability of the conclusions that are drawn. The analysis leads to the following observations:

- A majority of the respondents have their servers in a dynamic power mode. These modes result in a workload dependent power draw of these servers.
- All respondents apply some form of "high performance" setting by default.
- Many respondents apply conflicting settings on BIOS and OS level.
- Changing power management settings to more power saving modes results in approximately 10% energy savings in highly occupied server nodes. No adverse effects on performance were reported during the testing of these power saving modes.
- Changing Static high-performance settings to dynamic high performance does not necessarily lead to energy savings in a single server but can save energy of an entire cluster of machines.
- The best occupied servers still spend more than one third of their energy use on idle cycles, the worst occupied servers spend close to 99% of their energy on Idle.

Further discussions with the respondents about the reasons and barriers for not applying power saving modes lead to very consistent answers:

- Notwithstanding statement 4. there are still major concerns about performance losses when applying power saving. Even when this pilot returned no indication that performance loss occurs.

Given the caveat of a small study, the following conclusions can be drawn on the basis of the pilot outcomes:

- Employing the “power save setting“ contributes strongly to the goals for LEAP (i.e. improve energy efficiency of ICT in datacenters).
- The potential for energy savings from virtualization remains extensive. Pushing for higher levels can result in higher energy as well as financial savings than are currently targeted by the LEAP coalition.
- The data collected implies further research is needed. The concerns about performance impact indicate better understanding and research into power management is needed, including impact on application performance. Also, a more comprehensive statistical analysis of the use of power management features and average CPU loading is needed to draw strong conclusion about the general use of power management features and the energy potential.
- There is a pressing need for clear guidance and instruction from software and hardware providers, most ideally in unison, on how to best apply power management settings. This guidance must highlight the possible savings and explain when the standard power management can be tightened or must be relaxed.

Note that the use of “power management” and “virtualization” are measures within the framework of the “*Informatieplicht*” for data centers that are part of the “*Activiteitenbesluit*”.

CONTENTS

LEAP TRACK 1 'POWERMANAGEMENT' PILOT ANALYSIS 1

PREFACE 2

EXECUTIVE SUMMARY 2

1 INTRODUCTION 5

1.1 LEAP 5

1.2 DATACENTERS AND ENERGY 6

1.3 TARGET OF THE MEASUREMENTS 7

1.4 METHOD OF THE MEASUREMENTS 8

1.5 SERVER IDLE COEFFICIENT 10

1.6 DETERMINING PIDLE 11

1.7 SITUATIONS WHERE POWER MANAGEMENT IS INADVISABLE? 11

2 ACPI 12

3 DATA ANALYSIS 15

3.1 STATIC HIGH PERFORMANCE 15

3.2 DYNAMIC PERFORMANCE 17

3.3 PREDICTABLE DAILY VARIATIONS IN LOAD 21

3.4 UNDERLOADING UNCOVERED IN THE DATA SETS 23

3.5 SERVER DYNAMIC BEHAVIOR AND THE SERVER IDLE COEFFICIENT 24

4 QUALITATIVE ANALYSIS OF INTERVIEWS 32

5. WRAP UP 33

5.1 OBSERVATIONS 33

5.2 CONCLUSIONS 36

5.3 RECOMMENDATIONS 37

1 INTRODUCTION

1.1 LEAP

The Amsterdam Economic Board, NLdigital, Green IT Amsterdam, the Netherlands Enterprise Agency and Omgevingsdienst NZKG have started the Low Energy Acceleration Program (LEAP). Together with companies from the data center chain, knowledge institutions, the government and supported by the DDA, this coalition collaborates to realize energy savings with ICT within data centers with the aim to accelerate the transition to a sustainable digital economy.

The **objective** of LEAP is to offer inspiring perspectives for the introduction of (new) technologies and accelerating developments that could lead to energy reduction for ICT within data centers. We do this to provide a positive impulse for the future-proof growth of the sector.

In order to further the LEAP objectives, the Netherlands Enterprise Agency (https://www.rvo.nl) commissioned the pilot “LEAP track 1”

The **scope** of LEAP track 1 focuses on realizing energy savings with existing technology such as power management, utilizing energy efficient setting of servers without loss of performance. The solutions could also include virtualization (maximizing the capacity of the servers in relation to energy consumption) and using an objective measurement tool to structurally monitor and analyze energy consumption in relation to performance. The **ambition** is to work with leaders and front runners in the data center value chain to achieve energy savings of between 20% -40% by the end of 2022.

LEAP is a coalition of (currently) 20 parties who support the LEAP objectives to achieve energy savings with ICT within data centers. These are parties in the data center value chain:

- Data centers: Interxion, Iron Mountain
- Organizations with significant data traffic & customers of data centers: Booking.com, Deloitte, Municipality of Almere, Municipality of Amsterdam, KPN, NEP The Netherlands, OD NZKG, Rabobank, Royal Schiphol, Group SURFsara, VU University Amsterdam
- (Hardware) vendors: Dell Technologies, Hewlett Packard Enterprise, IBM, VMware and Red Hat
- Government: Municipalities of Amsterdam and Almere, EZ / Netherlands Enterprise Agency (RVO), ODNZKG
- Branch and network organizations such as NLdigital, GreenIT Amsterdam, DDA and Amsterdam Economic Board.

Two hypotheses were formulated at the start of the pilot:

1. **There is no direct (linear) relationship between the IT workload on a server and the energy consumption of this server.** Although this hypothesis is based on outcomes of several cases, it is important to investigate this hypothesis in a more structural manner. A new measure has been introduced for this purpose: the Server Idle Coefficient (SIC)
2. **Enabling power management functions on servers provides opportunities to save energy, without noticeably affecting the performance or availability of the server.**

This report contains an analysis of the data collected in both phase 1 and phase 2 of the LEAP Track 1 ‘Power management’ Pilot analysis. It highlights observations and draws conclusions based on these observations. Recommendations based on these conclusions are formulated at the end of this document.

1.2 DATACENTERS AND ENERGY

It is a universally accepted fact that datacenters are major energy consumers in today's economy. Estimates vary, but a figure of 2% of the national electricity production is often quoted. Because of this usage, the efficiency of datacenters has been a focus of attention for many years and improvements made by datacenter operators over the past decades have created a situation where further improvements on the facility infrastructure are not likely to result in significant energy savings for these datacenters. For instance, worldwide figures obtained by the uptime institute show a stalling of facility improvements (figure 1):

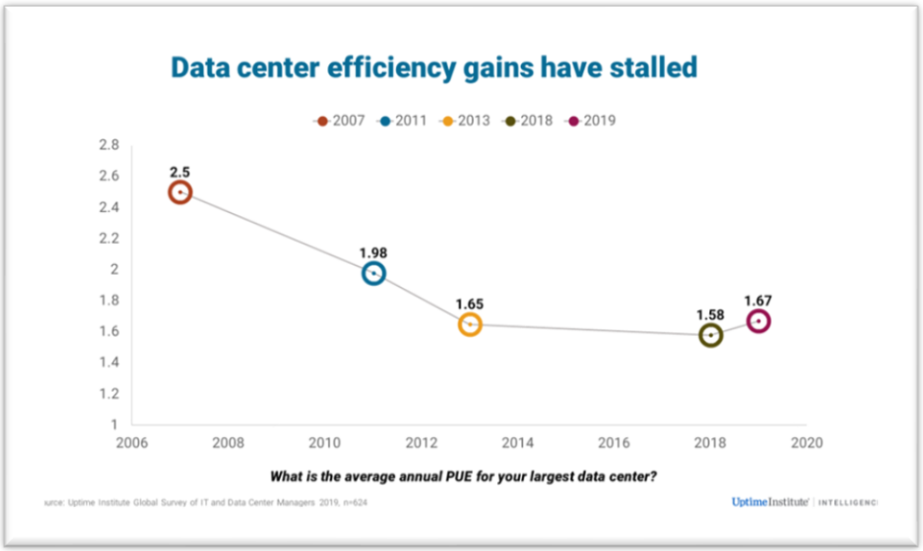


figure 1 : worldwide average PUE over the years

Power usage effectiveness (PUE) is a ratio that describes how efficiently a computer data center uses energy; specifically, how much energy is used by the computing equipment (in contrast to cooling and other overhead). Written differently;

Total Datacenter energy use = PUE X ICT energy use (1)

Formula 1 shows the importance of energy savings in ICT, the PUE acts as a multiplier, any kWh saved in ICT energy use results in a proportionally larger saving in the total energy used by a datacenter

In Amsterdam datacenters have a mandatory goal of a PUE of at most 1,3 but can reach PUE's down to 1,15 (This are "by design" numbers, while the graph shows the measured PUE) (source: Ruimtelijke Strategie Datacenters Routekaart 2030 voor de groei van datacenters in Nederland) The Dutch datacenters already outstrip global averages and further improvements are not likely to create a major improvement.

When improvements in PUE are no longer effective but significant improvements in total energy use are to be made, these improvements will have to come from advances in the ICT equipment with which these datacenters are filled. Such improvement in compute efficiency have been a part of equipment development for a number of decades, so much that these are coined into a "law" known as Koomeys law (source: Koomey post, https://www.koomey.com/post/153838038643)

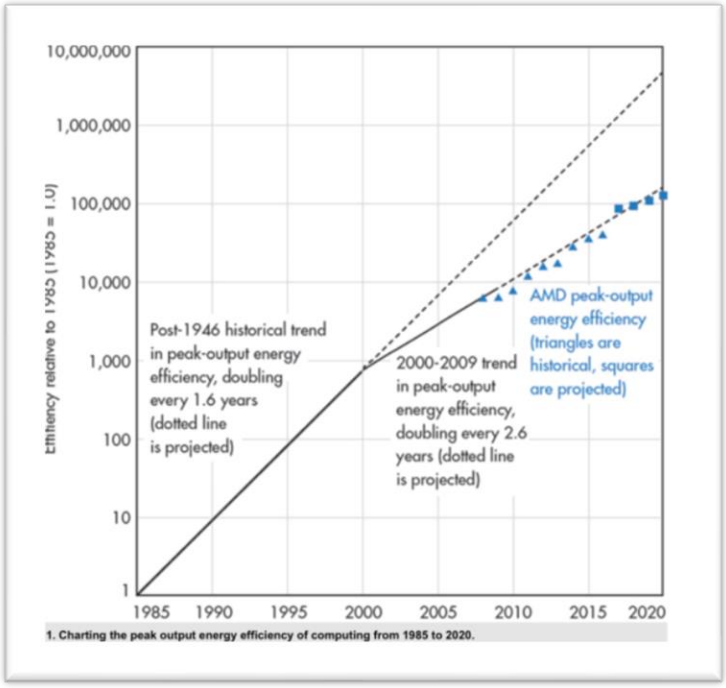


figure 2 : Koomey's law, advancements in compute efficiency

The improvements shown in figure 2 are however representative for maximum performance of a server. In practice, servers are rarely loaded to such levels and operate much closer to idle. The LEAP pilots will show data of various loadings recorded over a week that will illustrate a more realistic load than that used by Koomey.

1.3 TARGET OF THE MEASUREMENTS

In LEAP Track 1, we tried to determine the savings potential brought by power management functions available in modern ICT servers. To be able to analyse the hypothesis, the following measurements were taken during the pilots:

- Measuring Electrical power of servers
- Simultaneously measuring CPU utilization of these servers

The target of the pilot was to have these measurements done with two different settings for power management, first the baseline, a full week of measurements with the current settings, followed by a second week in which a more stringent level of power management was applied.

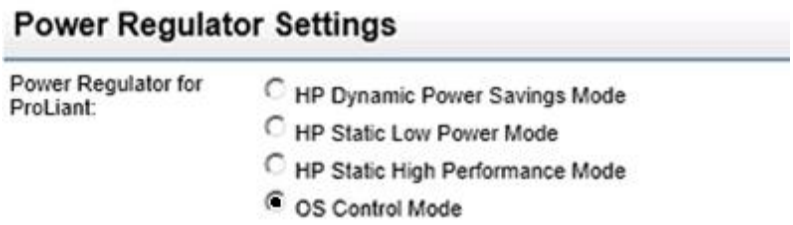
Such measurements are important in order to investigate whether power management can actually be utilized to realize energy savings without noticeable impact on performance. A Secondary goal was the testing of a novel metric: the Server Idle Coefficients and its sensitivity for the impact of power management.

1.4 METHOD OF THE MEASUREMENTS

For this pilot, a measurement protocol was distributed to the participants. This protocol consisted out of a few simple steps to record a baseline and an additional measurement for determining any effect of power management changes.

The first step was to record the current settings for power management in both BIOS and OS.
Although power management interfacing is standardized, different manufacturers offer different options for the BIOS settings and also use a different naming convention.

Examples of the options within the BIOS comes from a HP proliant:



and from Dell Poweredge servers where DBPM stands for demand-based power management.

Static MAX Performance	DBPM Disabled (BIOS will set P-State to MAX) Memory frequency = Maximum Performance Fan algorithm = Performance
OS Control	Enable OS DBPM Control (BIOS will expose all possible P states to OS) Memory frequency = Maximum Performance Fan algorithm = Power
Active Power Controller	Enable DellSystem DBPM (BIOS will not make all P states available to OS) Memory frequency = Maximum Performance Fan algorithm = Power
Custom	CPU Power and Performance Management: Maximum Performance | Minimum Power | OS DBPM | System DBPM Memory Power and Performance Management: Maximum Performance |1333Mhz |1067Mhz |800Mhz| Minimum Power Fan Algorithm Performance | Power

What is obvious that most systems use 3 different kinds of settings:

Static, where there is little to no relationship between server workload and power draw
Dynamic, CPU states controlled by the hardware
OS controlled, where CPU states are under the control of the master OS, the layer that does the hardware abstractions.

Within these operating systems similar options exist, for example with VMware ESX:

The OS setting only takes effect when the appropriate BIOS setting is applied.
As will be seen, the naming and additional comments for the power management settings are an important influence for the choice made by system administrators.
“high performance” by name alone seems a logical choice and as shown by the pilot, is the most used setting. Detailed examination of the actual processes behind these settings show that in many cases, the balanced mode can provide performance benefits and any degradation in performance has not been observed even in “low power” mode

After recording the current settings, a week of measurements is requested. Although shorter measurements periods will also yield results, A full week is preferred in order to be able to record workload variations that are the result of work patterns associated with business opening hours.

The measurement consists of 2 data points collected at least once every 15 minutes

Total power draw [Watt]
CPU utilization [%]

The powerdraw can be obtained from the systems management console, all modern servers supply this information to the system administrator.

CPU utilization is collected either from the master (host) operating system or from management software. CPU utilization is expressed as a percentage of available CPU capacity. CPU utilization is measured over a time interval where the percentage expresses the amount of clock cycles where instructions are processed as fraction of the total available clock cycles. A suitable interval must be selected to dampen fast variations, a rolling 20 seconds average is suggested common to many performance monitoring tools but the actual interval is left to the system administrator.

A short sample of such a measurement will look as follows:

Time stamp	CPU %	Power [W]
28/05/2020 12:16	24,16	364
28/05/2020 12:31	28,2	359
28/05/2020 12:46	53,57	408
28/05/2020 13:01	24,54	351
28/05/2020 13:16	24,43	356
28/05/2020 13:31	28,85	372
28/05/2020 13:46	35,7	377
28/05/2020 14:01	45,36	392
28/05/2020 14:16	29,22	367

1.5 SERVER IDLE COEFFICIENT

A new metric has been developed coined as the “Server Idle Coefficient” (SIC). The starting point for the development of this metric is a continuing search for an objective ICT efficiency metric. Efficiency metrics are defined as the amount of energy needed per unit of work. While the energy use in ICT systems is easily measured, the definition of a unit of work has never been agreed upon.

The new metric is based on the concept that “a unit of work” cannot be agreed upon, but, the opposite, a unit of idleness can be universally accepted. "Idle" is considered as period with no CPU load. For determining the idle coefficient, we measure the total energy use of a server and determine the energy spend in the idle state. The electrical power demand of the server in idle state is measured or otherwise determined. In the calculation this power draw is written as “P_{idle}”

In the LEAP pilot we measure power “P” and “CPU%” utilization we can calculate the SIC as

$$SIC = [E_{total} / E_{total} - E_{idle}] \quad (2)$$

Where the SIC varies from 1 to infinity (as is also the case with the well-known PUE).

Alternatively, a different representation of the SIC can be given as

$$SIC\% = [E_{idle} / E_{total}] \quad (3)$$

The SIC then varies between 0 - 100% and represents the fraction of energy used for the idle state.

A third representation has also been suggested:

$$SIC_{score} = 10 * (1- (E_{idle} / E_{total})) \quad (4)$$

In all of the calculations above, the energy used for idle in a period (n) is calculated by:

$$E_{idle}(n) = [100\% - CPU\% (n)] \text{ Pidle} * \text{interval length}(n) \quad (5)$$

$$\text{Total idle energy: } E_{idle} = \text{Sum } [E_{idle}(n)] \quad (6)$$

$$\text{Total energy: } E_{total} = \text{Sum } [P(n) * \text{interval length } (n)] \quad (7)$$

1.6 DETERMINING PIDLE

Determining the server power draw when the server is in Idle mode (Pidle) is essential for determining the SIC (see equation 5) but the determination of the Pidle is not trivial.
The ideal situation is to have a fully installed server, including the virtualization layers and OS installed but without any user programs running.
This situation is created in benchmark situations when determining the SPECpower benchmark. The total power draw is recorded with the system turned on, but without any programs running, yielding Pidle.

This ideal situation cannot be used when trying to determine the Pidle in active servers. These machines cannot be isolated and user programs cannot be stopped because of a measurement of idle power.

A series of other options exist for determining the idle power draw:

1) When a server has a static power setting, the active and idle power are identical. In this case, the equations for determining the server idle coefficient simplify and the SIC equals the average CPU idle percentage.
2) When a server has a dynamic power setting but shows a period in which CPU utilization is below 1%, the average power draw over this period can be considered a fair approximation of Pidle.
3) When a server has a dynamic power setting but is never completely idle, the linear extrapolation of the power vs CPU utilization curve towards 0% utilization will yield a acceptable value for Pidle
4) When only server power statistics are available as might be the case when limited or no access to the CPU statistics is granted, an average over the period of lowest recorded power use is assumed to be yield Pidle

Each of the methods has been used in the analysis of the LEAP pilot results.

1.7 SITUATIONS WHERE POWER MANAGEMENT IS INADVISABLE?

Power management is a collective name for technologies, so the question concerns the desired setting.

In a limited number of cases, the high-performance setting is preferable to the balanced or power save setting. High performance settings results in CPU cores not moving to higher C-states. This means that all CPU cores are always active. This is desirable when very consistent and fast response times are desired. Note that this is not about the total computing power of the server, but about the reaction speed to a command, even if a server itself has little CPU load.
Situations like this occur in High Performance Computing (HPC) in which, for example, RAM memory of multiple servers is combined over special networks. As well in the financial world, where AI is traded on the stock exchange, a millisecond delay can be too much. In these cases, the high-performance setting provides the desired functionality.

2 ACPI

In a computer, the Advanced Configuration and Power Interface (ACPI) provides an open standard that operating systems can use to discover and configure computer hardware components, to perform power management by (for example) putting unused components to sleep, and to perform status monitoring. First released in **December 1996**, ACPI brings the power management under the control of the operating system, as opposed to the previous BIOS-centric system that relied on platform-specific firmware to determine power management and configuration policies. The specification is central to the Operating System-directed configuration and Power Management (OSPM) system, an implementation for ACPI which removes device management responsibilities from legacy firmware interfaces via a UI.

Intel, Microsoft and Toshiba originally developed the standard, while HP, Huawei and Phoenix also participated later. In October 2013, ACPI Special Interest Group (ACPI SIG), the original developers of the ACPI standard, agreed to transfer all assets to the UEFI Forum, in which all future development will take place. The UEFI Forum published the latest version of the standard, "Revision 6.3", in end of January 2019.

Simply put, any server in operation today has the ability to match its electrical power usage to its ICT workload in some degree. The control of the dynamic range is either in the hardware itself (through BIOS settings) or in the Operating System (OS) running on the hardware. The OS is meant here in its broader term: VMware ESX or Microsoft Hyper V are just as valid as any Windows, Linux or Unix OS.

It is important to note that these to control mechanism are presented to the system administrators simultaneously, the correct modus for OS controlled operations would need a BIOS setting of “OS-controlled” followed with an appropriate setting within the OS. It is likely, but unconfirmed at this point that any other BIOS setting will override OS settings, but further research would need to show the effect of conflicting settings in BIOS and OS.

Power management has several steps:

- HP = High Performance - this means that little energy is saved when the server's CPU is idle. In many cases adjustments will still be made to the clock frequencies. These adjustments fall under the so-called ACPI-P states. These adjustments happen if a CPU is not idle, but underloaded.
- additional power management steps - many servers have multiple power management settings. These can be specific per brand and type of server. They serve to achieve more and more energy savings, as the need for CPU capacity decreases further and / or one or more core (s) are switched off (deeper). For example, CPU status:
    - C0 = active
    - C1 = least aggressive form of downshift with idle
      Wake-up time of a switched core is roughly 0.5 microsecond.
    - C6 = heaviest C-state, CPU has no power at all
      Wake-up time from C6 = roughly 40 microseconds

If CPU "runs" at 3.3 GHz, then wake-up from C-states to C0 is within 1650 - 13,200 clock cycles

In order to put these possible delays into perspective, the first thing to realize is that for a CPU to be put into a C6 state, this particular CPU must not have been used for a considerable time. ACPI-C states apply to idle CPU’s only, the quoted wakeup latency is a hit that only happens once, when an idle CPU needs to be added to the pool of active CPU’s.

The second element is to look at various other delays that can occasionally happen during a computation, the response time of a hard disk is in the range of 10 ms but also network traffic can introduce ms delays. Even without any handling delays (send/receive), the round-trip time over 100 m of optical fibre alone is 1 microsecond. It is fair to conclude that it would be impossible for an end-user to detect an additional 40 microseconds delay in the response time of an application.

The working of ACPI states (power management) seems particularly useful when we take into account the workload profile such as published by the Amsterdam internet exchange (AMS-IX)

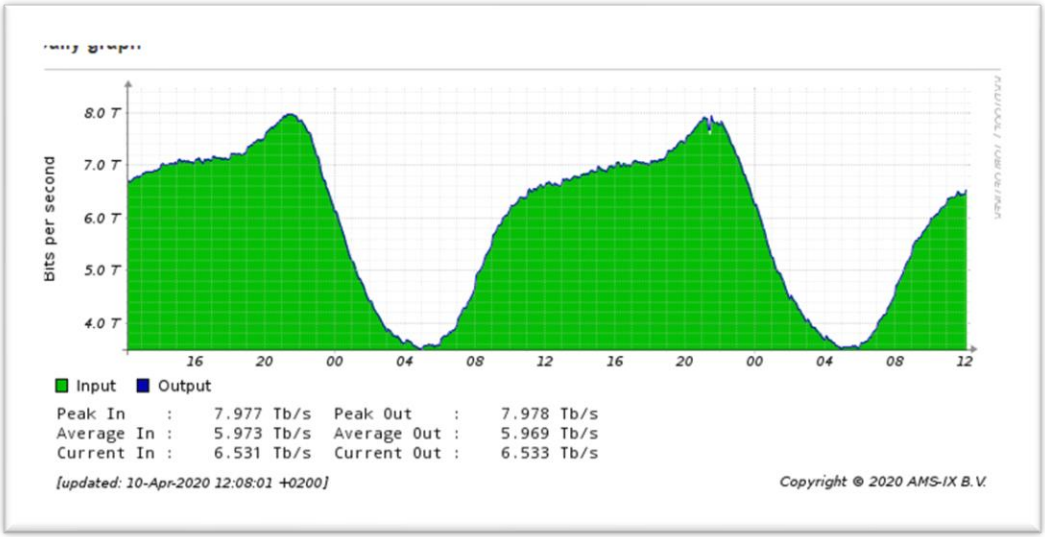


figure 3 : AMS-IX daily traffic

These graphs show network traffic over two 24-hour periods and demonstrate the huge difference in internet traffic over a day. One must assume that a similar variation in server CPU load should accompany the dramatic variations in network traffic. Through the use of the power management states, servers can lower their energy use when the workload decreases.

Two types of server load are distinguishable in the LEAP pilot, namely machine to machine type applications and Machine to end-user applications.

In general, one can observe the CPU load of servers running machine to user applications to have a recurring pattern of high load when users are active and low load when not. In commercial environments, these active times often correspond with office hours. The most obvious machine to end-user type application shown in the data below is that of the “virtual desktop” application such as Citrix (see figure 10). Users login in the morning, work and log out during the evenings. Servers running only this type of workload are essentially idle for 120 out of 168 hours per week and would benefit greatly from the highest possible settings for power management.

Machine to Machine type applications are not directly connected to any end-user activity. Prominent in the pilot are servers that run monitoring applications. These applications monitor the health of networks, systems and applications and do so through regular polling. Such applications do not exhibit idle periods and in general result in a very constant workload. The predictability of these workloads makes it possible to optimize workload placement over the available hardware resulting in high average CPU loading with

very little variation. (see figure 9). Such systems are unlikely to have CPU's switching between idle and active states (due to the constant workload) but can still benefit from power management because not all computational power configured in these machines is needed to perform their tasks.

3 DATA ANALYSIS

A total of 13 companies indicated willingness to provide both baseline data and changing of power management settings during the kickoff meeting of 12th December 2019. After an extension of the measurement period, a total of 9 parties provided data by September 2020. All contributions have been anonymized. Several server systems, selected by the owners of the systems, have been monitored and their CPU usage as well as power draw have been recorded. In the following paragraphs specific results from the monitored devices are highlighted and analyzed. These specific data sets were chosen on the basis of the situation/setting/effect that was highlighted by the particular data set. Conclusions drawn on the basis of the data are then made in chapter 5.

3.1 STATIC HIGH PERFORMANCE

Out of the datasets obtained during the measurement period, one set shows a company that consistently applies **static high-performance** settings on their HP blade infrastructure.
The effect of applying a static setting on the hardware level is demonstrated in figure 4.
The measurements were taken at 15 minutes interval, each graph contains a week worth of data from a single server.

Servertype	HP BL460C Gen8 (2016)
power management	
hardware (BIOS)	**STATIC High Performance**
OS	High Performance
CPU Type	Nr.
Intel(R) Xeon(R) CPU E5-2680 0 @ 2.70GHz	1
Intel(R) Xeon(R) CPU E5-2680 0 @ 2.70GHz	2
Operating system	VMWare

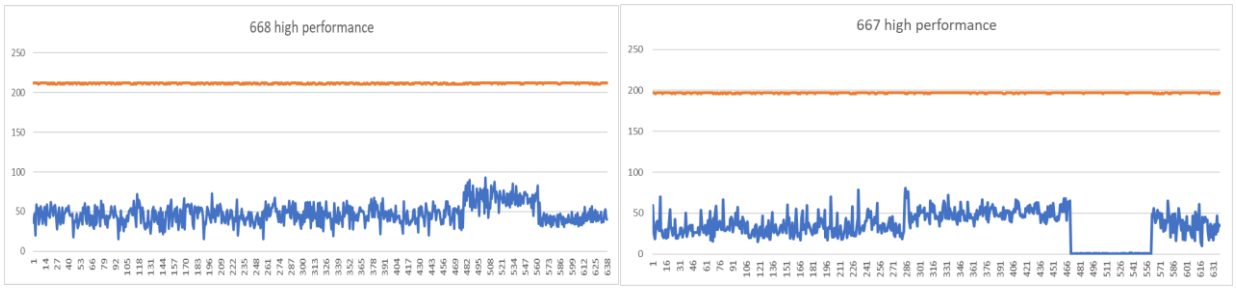


Figure 4 : two servers in static high performance mode. In these graphs, the vertical axis have a double meaning. The blue line shows the CPU load as a percentage from 0% (idle) to 100% (fully loaded), the orange line shows the electrical power draw P in Watt. The horizontal axis just shows the number of the measurement interval.

As can be seen, there is no variation reported in power draw. The graphs were selected specifically to show the extend of the effect. The CPU load of zero in server 667 during a 24-hour period has been confirmed as real, the VMware cluster correctly absorbed the workload from this node during the period (see figure 4, graph on server 668).

It is very interesting to see the effect of changing the setting to an OS-controlled mode. Two preliminary observations have to be made here;

1. Switching the power management control from hardware level to OS control did require a reboot of the server. For this particular configuration this was cited as one of the reasons why the change was not effected on other platforms.
2. The choice for these servers was also motivated by the necessary reboot, coupled to the lack of knowledge about any influence on application performance. The system under observation is for internal use by the systems management division. It runs machine to machine type applications which in turn explains why there is no consistent daily workload variation visible.

power management	
hardware (BIOS)	**OS-controlled**
OS (VMware)	Balanced Performance

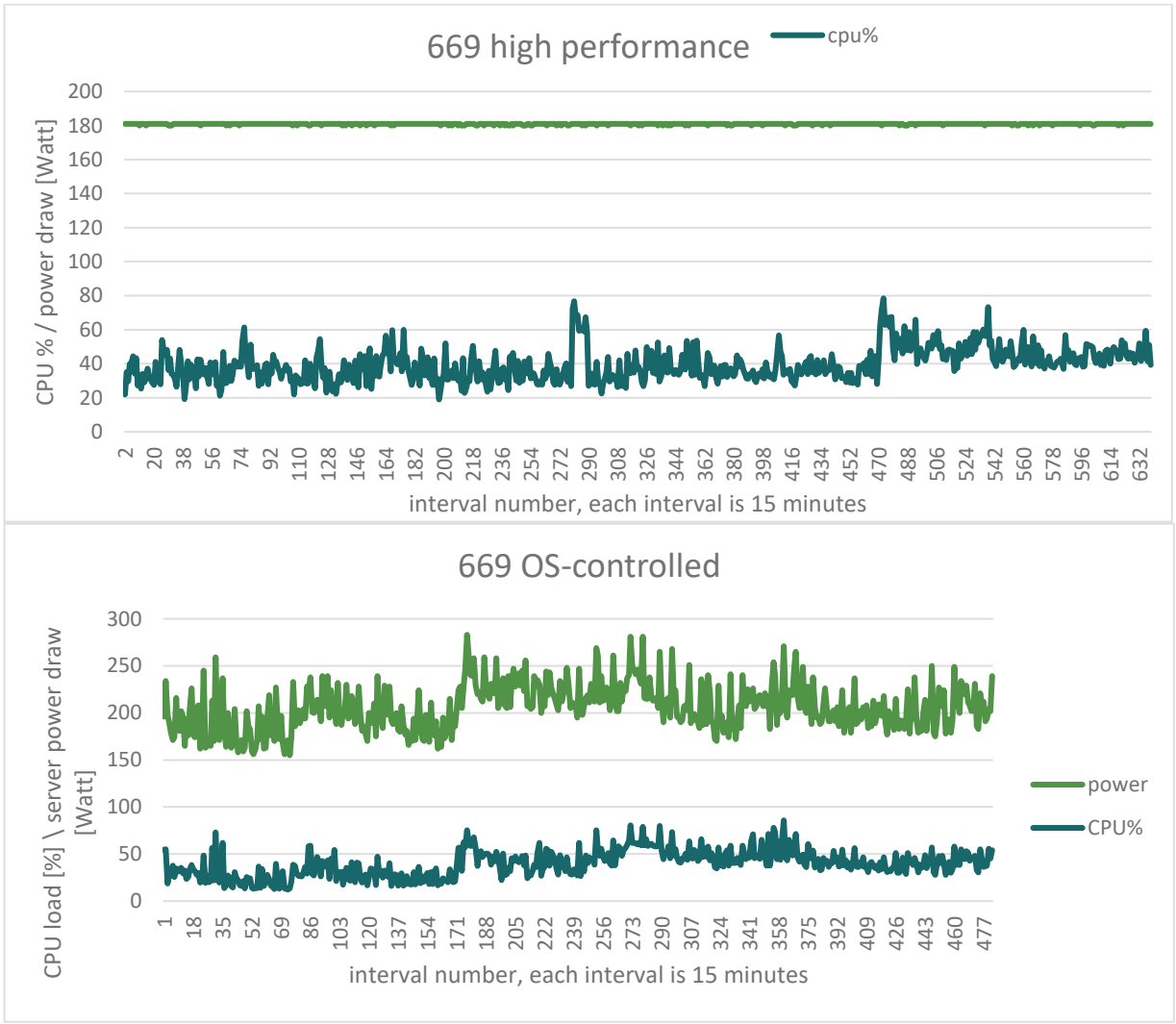


figure 5: single server configured with static (top) and dynamic (bottom) power. In these graphs, the vertical axis have a double meaning. the CPU load is shown as a percentage from 0% (idle) to 100% (fully loaded), the electrical power draw P of the server is shown in Watt. The horizontal axis just shows the number of the measurement interval.

Aside from the graphs, averages over the measurement period can be calculated:

Weekly averages STATIC High performance:
Power **180,9**
CPU% **39,3**

Weekly averages DYNAMIC (OS controlled):
Power **205,2**
CPU% **39,6**

As can be seen the results are unexpected. The static high-performance results in a lower average power draw than the OS-controlled mode. Although it is impossible to test what is really happening in this situation, it is suspected that the “static high-performance mode” is achieved by disallowing high CPU Pstates. These high frequency states sometimes referred to as turbo modes are power hungry but do deliver higher performance. It is visible in the data that during low utilization the power draw drops below the 180W associated with the static high-performance mode. But at CPU load of 40% and more, the power draw of the CPU rises considerably pushing the system to a power draw of 250 Watts.

It is very likely that the application performance is much higher during the OS-controlled period than under static high performance. But there is no data to substantiate this claim.
What is confirmed is, that power draw under idle conditions is much lower in the dynamic mode than in the static mode. The fact that the total energy use in this particular case rises can be attributed to the consistent high application load. The particular setting will most likely result in energy savings in less heavily occupied servers.

3.2 DYNAMIC PERFORMANCE

The most commonly encountered setting for the servers that were monitored was a variation on the theme of dynamic high performance. In this mode, there is a definite response of the server to decreasing and increasing load.

Correctly applied on a large VMware cluster of one of the respondents was the OS-controlled mode in two different settings:
BIOS: OS-controlled

HPE BL460 Gen9
2 x Intel(R) Xeon(R) CPU E5-2697A v4 @ 2.60GHz
+/- 60 vm’s per node in the clusrter 20 nodes
Hypervisor: VMware ESXi 6.5.0 build-13635690
Measurement 1: high performance
8-1-2020 to 17-1-2020 10:43
Measurement 2: balanced
17-1-2020 10:43 to 20-1-2020

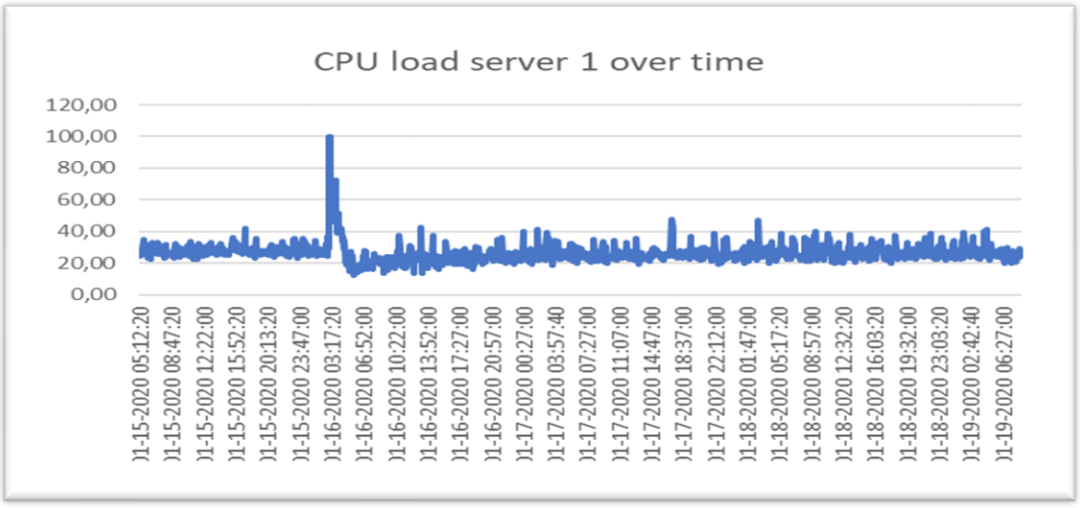


Figure 6: server 1 CPU load over time

The server used for the LEAP pilot is used for company internal use only. It is a highly virtualized environment with a very high VM count per physical node. What is apparent is that the average total CPU load from the 89 VM's do not show a relationship with the time of day. Load is essentially constant. The peak (100%) occurs at 3 AM on the 16th, cause is unknown.

A useful way of displaying the measurements is by creating a graph that shows the relationship of the CPU loading and the server power draw. A clear relationship is considered desired behavior and as shown in figure 7, even in the high-performance mode, this relationship is apparent.

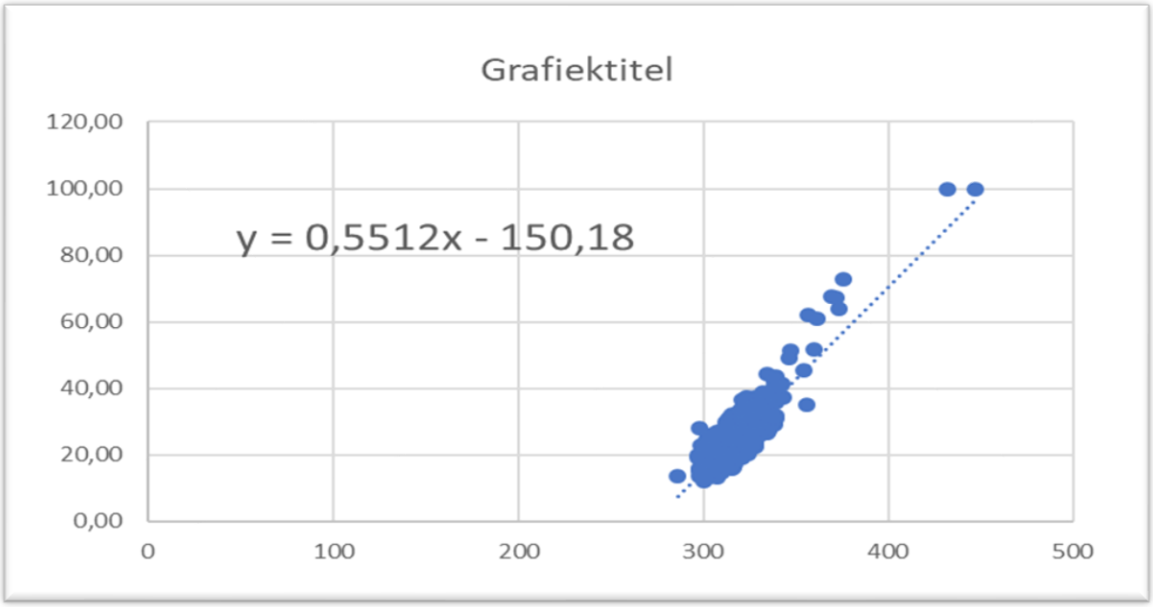


Figure 7: server 1 CPU load vs power, high performance. In this graph, the vertical axis shows the CPU load of the server and the horizontal axis shows the power that was measured. Each point corresponds to a measurement interval where both CPU loading and power data were collected. (for example, 20% load, 300 Watt)

The dashed line is the best linear fit to these measured points. Extrapolation of this line to a CPU load of zero gives us a calculated power draw when idle (P_{idle}) of 273 Watt. The error margin is substantial because of the high number of measurements that fall between 10% and 40% CPU utilization. The use of this extrapolation method is further discussed in paragraph 3.5.

A similar graph can be created of the measurements made under the “power save” setting:

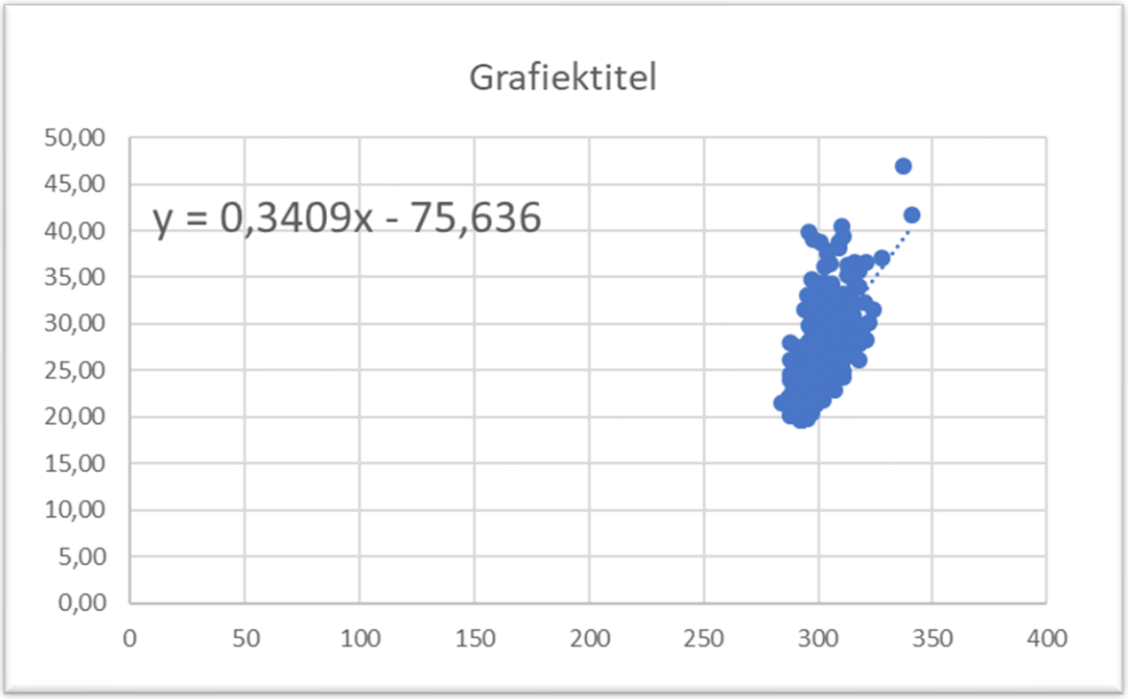


Figure 8: server 1, power save setting. In this graph, the vertical axis shows the CPU load of the server and the horizontal axis shows the power that was measured. Each point corresponds to a measurement interval where both CPU loading and power data were collected. (for example, 25% load, 300 Watt)

The extrapolation of this dashed line results in an idle usage of the blade of 222 Watt. The error margin is again substantial, but the influence from the power save setting is clear. Even under constant substantial load, power efficiency shows a substantially lower power draw and will lower energy use.

A straightforward averaging of the data from the two measurement periods confirms this:

Server measured with setting “high performance”
Average power: 321,4 Watt
Average CPU 26,97 %

The same server measured with setting “balanced”
Average power: 300,4 Watt
Average CPU 26,7 %

The effect of the change is not limited to a single cluster node, taking the average load and average power draw of all 20 nodes in the cluster during

Cluster measured with setting “high performance”
Average power: 271 Watt/node
Average CPU 18 % / node

The same server measured with setting “balanced”:
Average power: 252 Watt/node
Average CPU 17 %

There are two important observations that can be derived from the data:

1. Even when in OS controlled, High performance mode, the server still exhibits dynamic behavior. This is a clear break with the situation shown in paragraph 3.1, where under “static high performance” there is no dynamic behavior detectable.
2. Even under very high load conditions, enforcing the “balanced” mode in VMware still results in a 7% energy saving. No adverse effects on application performance were reported during the power save period.

The third respondent that did apply changes to power management settings used hardware control only.
The servers in question were switched from week 1, BIOS setting: Efficiency - Favor Performance to Week 2, BIOS setting: Minimal power
The servers measured however show almost no variation in load and with the exception of questionable values almost no variation in power draw.
The observed constant CPU load is in line with the function of the server. The servers is again for internal use only and the application running is a monitoring application, collecting data from other servers for management purposes.

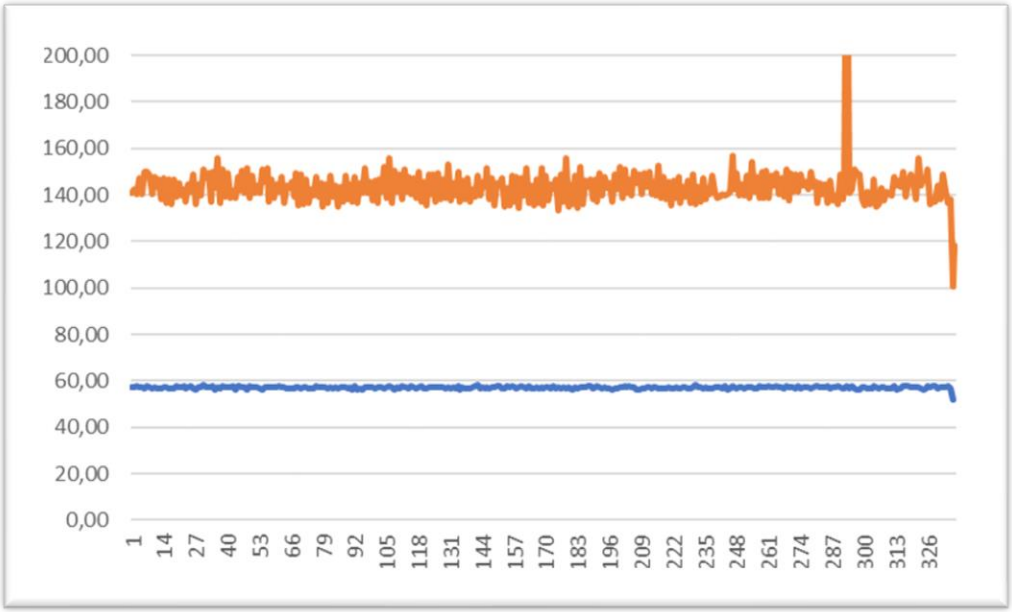


figure 9: single server configured with “favor performance”. In this graph, the vertical axis hase a double meaning. The blue line shows the CPU Idle as a percentage from 0% (fully loaded) to 100% (idle), the orange line shows the electrical power draw P in Watt. The horizontal axis just shows the number of the measurement interval.

The most logical method for studying the effect of the change in power setting is to take straight forward averages of the data, the following variables were monitored:

Week 1:

Power	CPU			
Watts	System	User	Idle	IO Wait
143,46	19,42	21,68	57,03	1,87

Week 2:

Power	CPU			
Watts	System	User	Idle	IO Wait
126,04	19,38	21,64	57,11	1,88

It was clearly indicated that adjusting power management had no detrimental effect on performance and even at a near constant utilization of 40%, power management has effect. It is probable that some CPU cores remain untouched by the application (40% utilization) and are therefore moved to a higher C-state.

Changing power setting results in a 19.4 Watts saving, = 13%

3.3 PREDICTABLE DAILY VARIATIONS IN LOAD

Some of the servers that were part of the pilot showed a very clear behavior that corresponds with the expected day and night variation as shown by the AMS-IX daily load graphs.

The clearest example of predictable load variations come from a Citrix VDI server of one of the participants. The peaks in CPU load correspond perfectly with the peaks in power draw. What is unexpected is that the server is reportedly in the so called “static high performance” mode. Why the server behaves in such a near perfect dynamic mode is unknown. It is possible that the setting shown has not been correctly recorded or if correct, has, for some reason, not taken effect. Another possibility is that, much like what is seen with other participants, the previously truly static “high performance” mode is still dynamic. The underlying details of each mode were not recorded and might differ between servers.

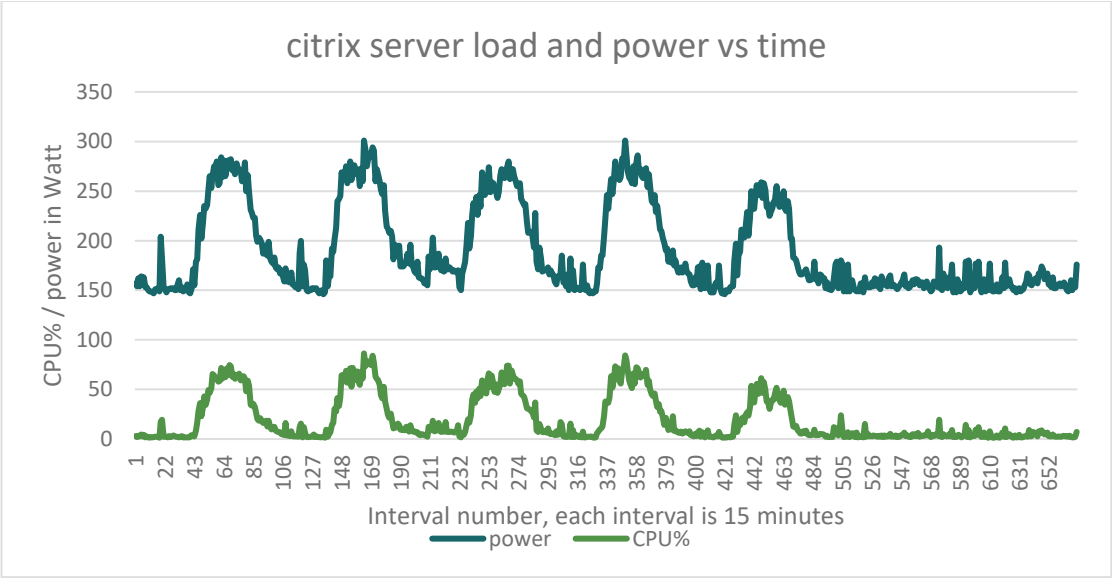


figure 10: Citrix server configured as "high performance". In this graph, the vertical axis has a double meaning. The green line shows the CPU usage as a percentage from 0% (idle) to 100% (fully loaded), the blue line shows the electrical power draw P in Watt. The horizontal axis just shows the number of the measurement interval.

Shown in figure 10 is a full week of measurements, with the first peak on Monday and the last on Friday slightly lower than the other workdays with a noticeable sharper drop of as people go home for the weekend. The data contains a number of datapoints for which the CPU load is actually 0.
This allows us to calculate the Server Idle Coefficient without resorting to linear extrapolations

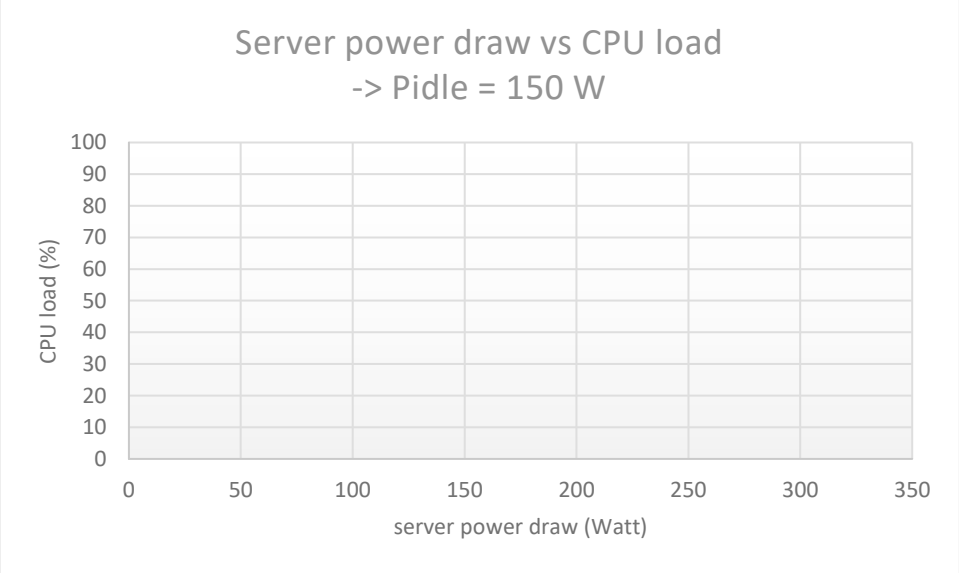


Figure 11 : Power vs CPU load

Figure 10 and 11 show both expected and desirable behavior for a user-oriented service. Clear load variations which are related to working hours, clear power variations aligned with the load variations. There is no measurement done with an adjusted power management level.

Summing the data over the measurement periods (see equation 5-7) gives us:
Idle energy 20,2 kWh
Total energy 31,9 kWh
Average CPU load 19,9% (CPU IDLE 80,1%)

Resulting in a
SIC% = 63,3%
Or in another representation:
SIC = 2,7

It is clear that any server that is so tied to office hours spends nearly ¾ of its time in very low load conditions. (weekends and nights) It is due to the dynamic power behavior that the SIC drops to 63% from the CPU idle of 80%.
Still, over 60% of the energy is spend in idle, it is an expectation that Increasing power management levels in this case would lower idle energy, thus decreasing overall energy spend and improving the SIC.

In the discussion with the party responsible for the infrastructure it was suggested that an additional step in power management was being considered. S-states. System states allow a controlling system to switch of entire systems in a cluster when the load reaches a certain defined lower limit. Any residual workload can be moved to a surviving cluster node. Deploying these states would lower the idle energy by approximately 10 kWh per week, this of course also drops total power by 10 kWh, the resulting SIC would be in the order of 2 (SIC% = 50%)

3.4 Underloading uncovered in the data Sets

Another observation from the data supplied was that there is a noticeable difference between the systems that have been optimized for maximum virtual machine placements and those that have either dedicated purposes or are more directly involved in providing services to end users.

Underloading of servers or stated differently over sizing the infrastructure needed to perform a certain service is still the costliest misuse of resources. Not only does such a configuration incur high operational costs (maintenance, licenses and energy), it is also costly from a CAPEX standpoint since large investments have been made in datacenter and servers that could have been avoided.

Examples of such oversizing are apparent in many of the provided datasets, the graph below is just one example of such a configuration. The system in question is configured in the state of “maximum

performance” which corresponds with static high-performance BIOS setting.

Power management:	
(BIOS)*	Max performance
CPU Type	RAM
2 x silver 4110	128 GB

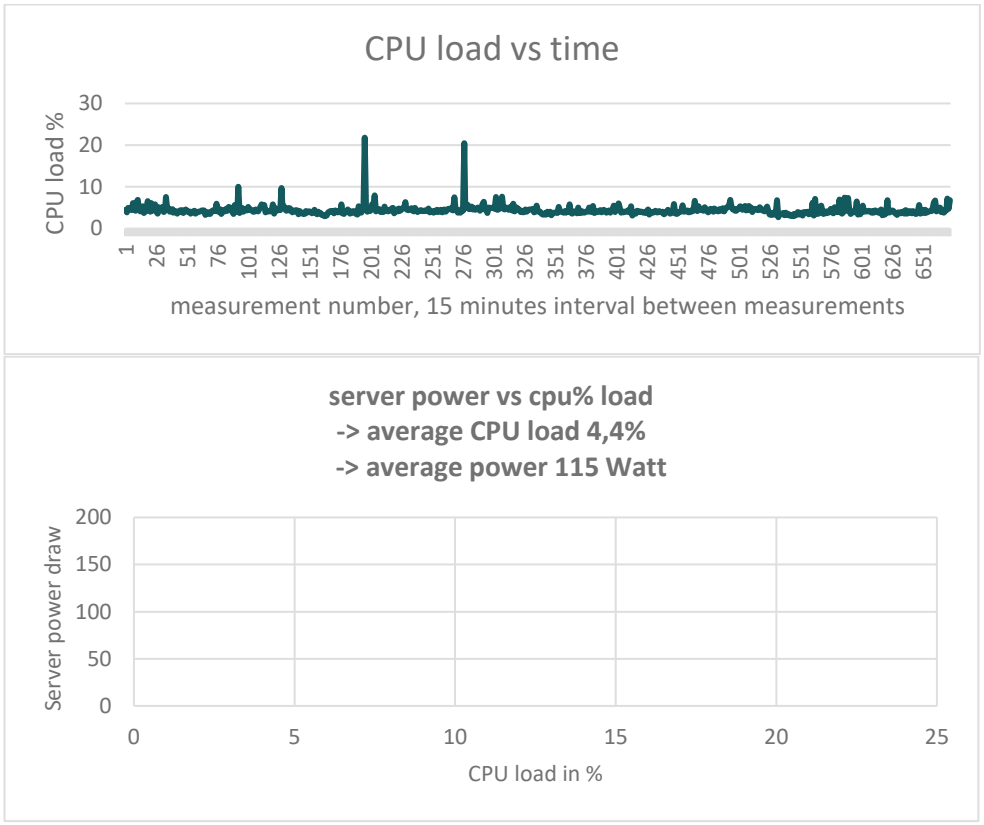


figure 12: underloaded server configured as "static high performance".

In figure 12, two graphs are shown, the CPU load vs time (above) and the power vs CPU load (bottom) analyzing CPU% (with a fully loaded system is at 100%) We observe that during the measurement period, only 3 instances of load over 10% were recorded. The Average CPU load = 4,4%, Power = 115W

The power management setting results in a virtually zero dynamic range, in such conditions, P_{idle} = P and the SIC% equals the CPU idle %,

SIC%= 95,6% or

SIC = 22,7 in the PUE like representation

As stated, severe and structural underloading of CPU's is a common occurrence. The current document does not show every data set obtained. Several systems in the data set show an average CPU load below 4% and even peak load in these systems rarely reach 10%. Such systems will benefit from the power management setting, but much energy and capex can be saved through consolidation of these workloads in better utilized (shared) environments

3.5 SERVER DYNAMIC BEHAVIOR AND THE SERVER IDLE COEFFICIENT

The range of dynamic behavior in various data sets is significant, but rarely do servers consistently reach 0% utilization as is the case with the Citrix server in paragraph 3.3. In order to find the power draw in idle condition, P_{idle} , which is necessary for calculating the Server Idle Coefficient, the power vs CPU utilization data was fitted to a linear function. The data showed a surprisingly good fit to such a linear approximation over a large range of CPU utilization numbers in all data sets obtained during this study.

From figure 11, the Citrix server that goes through the entire range of utilization numbers from 0% to 85% and from data published in various benchmarks available online, we know that such a linear approximation does not fit well around the utilization extremes (0-10% and 90-100% CPU utilization) but the data shows this approximation to be very close to observed behavior in the intermediate utilization range.

The “trend line” found can be used to extrapolate towards 0% utilization. We have used this method in order to determine the idle power draw, P_{idle} , from the data supplied. With the use of formula (5) we can calculate E_{idle} and from there the SIC/SIC%/SIC_{score}

Below in figure 13 we show such an extrapolation, the dashed line corresponds to the formula:

$$y = 0{,}8481x - 130{,}02 \quad (8)$$

where y represents CPU utilization (%) and x Power (Watt). Setting y= 0 yields x = 153W, the value for P_{idle}.

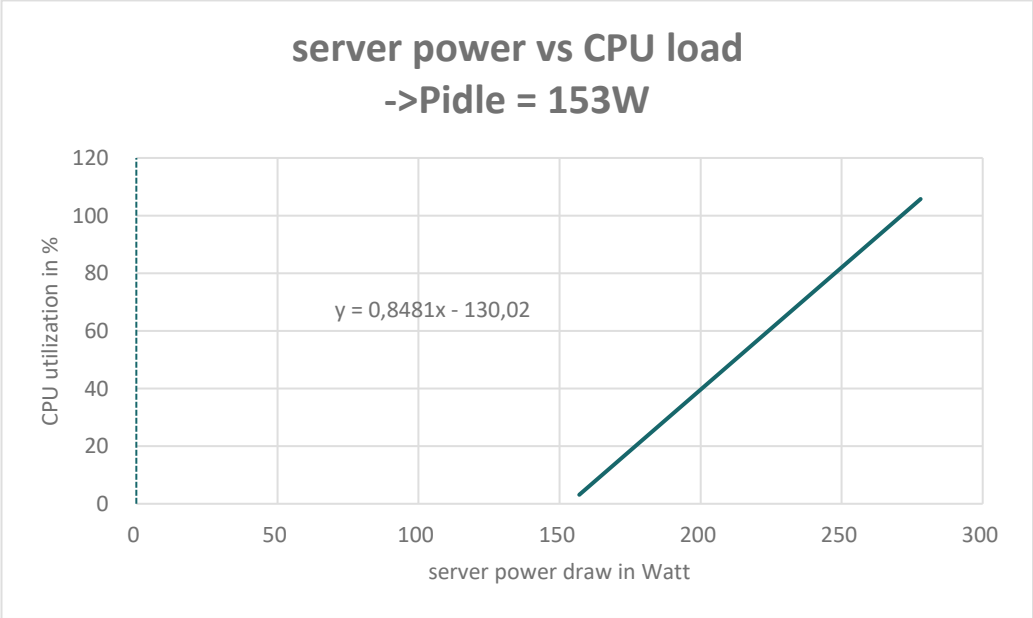


figure 13: CPU load vs power draw. Server in OS-Control mode running VMware with a balanced profile

Energy Idle	Energy total
17,96 kWh	31,25 kWh

SIC% = 57%
SIC = 2,3

In this particular situation, there was no corresponding measurement with a more stringent power management setting.

When we analyze the servers who's data is shown in the previous paragraphs we can observe the sensitivity of the SIC to changes in power management settings;

Switching from static high performance to dynamic balanced performance (figure 5 paragraph 3.1)

We can calculate the SIC for the server 669:

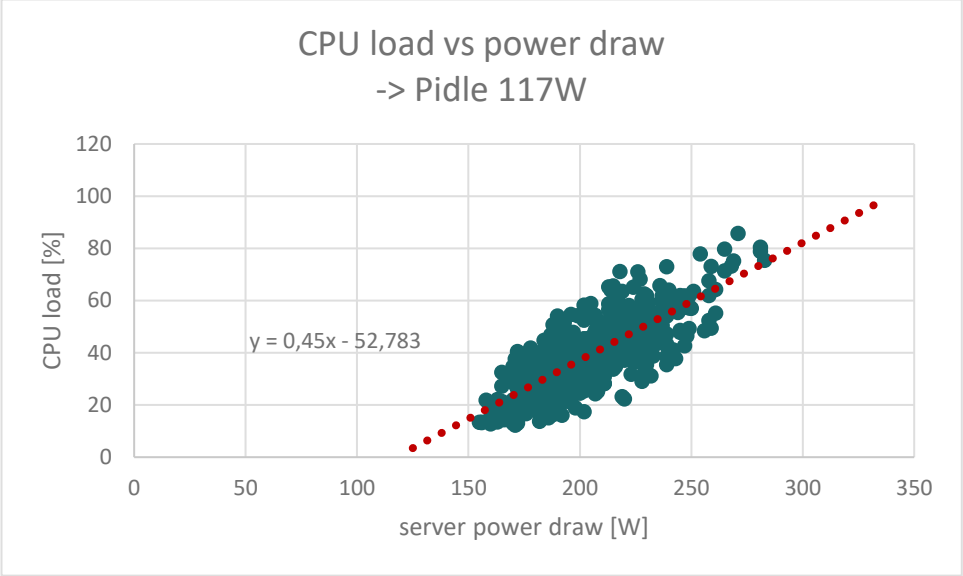


figure 14: cpu load vs power for server 669 in balanced mode, Pidle from extrapolation 117 Watt

From the data on server 669 we can derive the following numbers:

Total energy over the period: 24,5 kWh

Total energy spend in Idle: 8,43 kWh

Average CPU time in Idle: 60,4%

Using the appropriate formulas, we calculate the SIC% and SIC in balanced mode as:

SIC% = 8,43/24,5 = 34,4%

SIC = 1,5

As is visible in figure 5, when the system is in static high performance, Pidle = P and consequently

SIC% = CPU idle% = 60,4%

SIC = 2,53

These calculations give the indication that the SIC is a potent indicator and sensitive to a dramatic change in power management. The calculations also indicate that changing to a dynamic power mode might not result in absolute savings but does result in a shift to a more useful energy use, although the effect on application performance should be part of any following research into the topic.

The systems mentioned in paragraph 3.2 were subjected to a much smaller shift in power management settings. Namely from an already dynamic mode to a more stringent regime in which deeper sleep states are accessible. Server 1, figure 7 and 8 resulted in the following numbers:

Server measured with setting “dynamic performance”
Average power: 321,4 Watt
Average CPU 26,97 %
SIC% = 64% (SIC= 2,8)

The same server measured with setting “dynamic power efficient”:
Average power: 300,4 Watt
Average CPU 26,7 %
SIC% = 64% (SIC= 2,8)

In this specific case, the SIC turns out to be completely insensitive to the change in power management. Again, further research is needed but it appears that for this particular mode, both the maximum power draw (at 100% load) as well as power draw under idle conditions is lowered. The lowering of the idle power can be associated with the use of deep C-states, these states are only used in idle CPU cores and will not influence performance. The lowering of the maximum power draw however might be attributed to the abolition of high-performance P-States (turbo modes) Which would impact the maximum performance of the server. In the intermediate regime (up to 80-90% CPU load) where there is still ample headroom, the effect on application performance from the absence of turbo modes can be imperceivably small but again, these observations are a clear indicator of the need for a more comprehensive study that includes application performance measurements.

The data from the increased power management setting in the hardware shown in paragraph 3.2 figure 9 is not usable for a SIC calculation. The variation is CPU loading in this particular situation is so small that there is no way to observe any power draw variation over time. Hence, P_{idle} cannot be determined.

Lastly, the SIC can be used to analyze server behavior where CPU statistics are either absent or unusable. In such a case recorded at one of the participants, the CPU stats are very likely to be recorded in to short an interval. Since the system under study is highly underutilized, the CPU statistics show erratic behavior comprised of mostly 0% utilization. Interestingly, the server powerdraw shows a much more smoothed out pattern that reflects the expected load on the server. The server in question is used for office productivity and as such is virtually idle at night and as the measurements show very lightly loaded during the day.

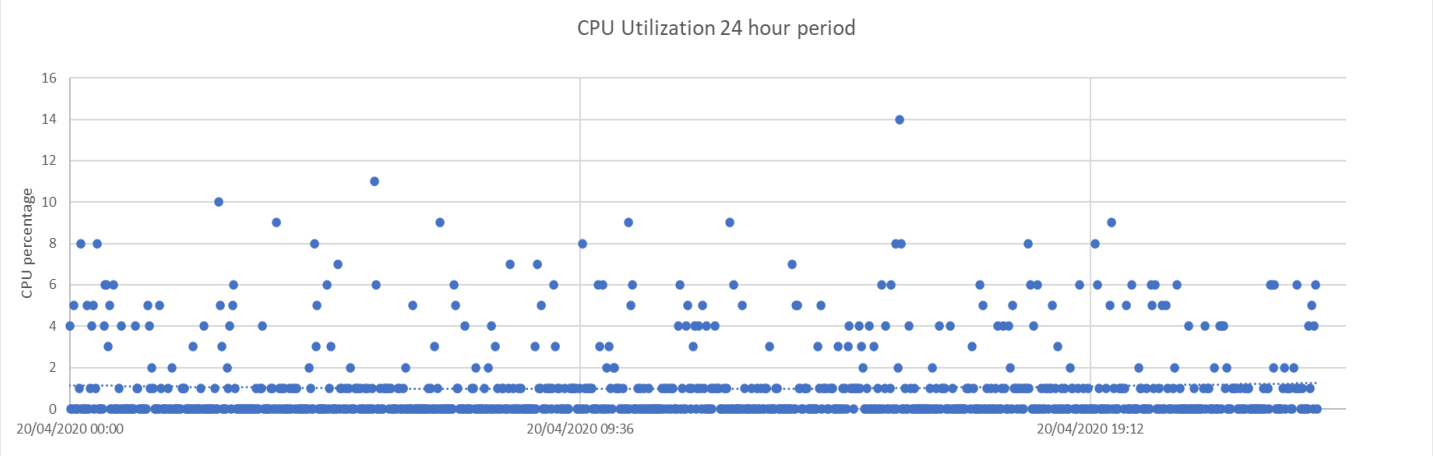


Figure 15: 24 hours recording of CPU usage

With a single peak of 14%, the average of the measurements is 1,0% utilization.

The system in question was actually measured in four different states:

The first two measurements were done to validate the newly formed expectation that the so called High performance state is no longer static, it also served as a comparison between hardware controlled and OS controlled High performance.

Shown if figure 16 are these first 2 measurements, both in high performance setting, Blue is the hardware control (BIOS in high performance) Orange in with OS control, meaning that the system was rebooted with BIOS power setting “OS controlled” and OS power setting “high performance”.

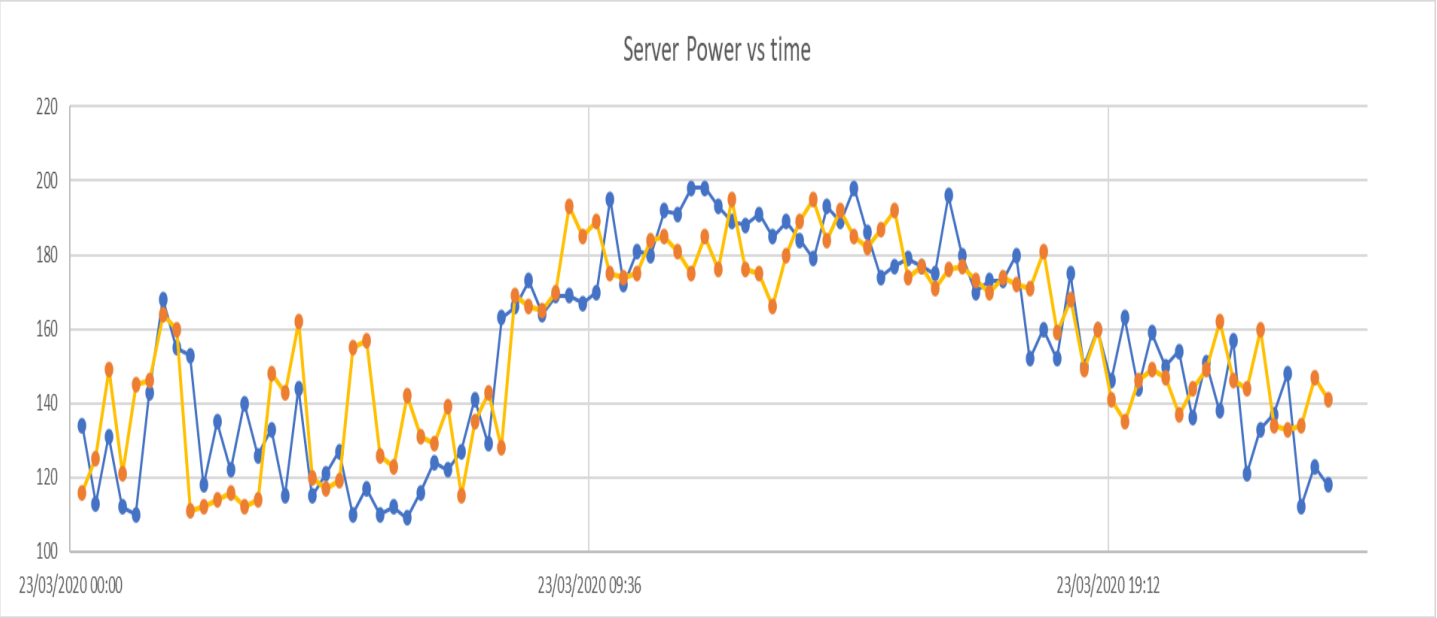


Figure 16: Power draw of server, Blue line hardware controlled, orange line OS controlled

It is discernable in figure 16 that between the hours of midnight and the start of the workday at 7:30 AM, the server under study is mostly idle. From this period, we can estimate P_{idle} as 110W

The energy usage of this system over the 24-hour period can be calculated as the area under the line. This area in both cases comes out at :

E_{total} = 3,72 kWh

Since CPU loading of the system is very low, we estimate the idle energy use as 110 W X 24 hours

E_{idle} = 2,64 kWh

Consequently, SIC = 71%

Given the fact that the average CPU load is very low (99% Idle), we can see the positive effect of the dynamic power behavior.

After discussing these results with the participant in question, a change to the OS power management setting was effectuated. The third measurement was conducted with the OS power management set at “balanced performance” and a fourth measurement with OS setting “low power”

During these measurement periods there has not been any mentioning of changes in application performance. Given the low system utilization and the earlier observation that higher CPU turbo modes are accessible with the balanced power setting, the expectation would be that perceived performance would be better with the “balanced” setting but independent application performance measurements have not been carried out.

The effect on the power draw of the system is however dramatic. The third measurement with the "balanced" setting is plotted together with the previous measurements, yielding the grey line:

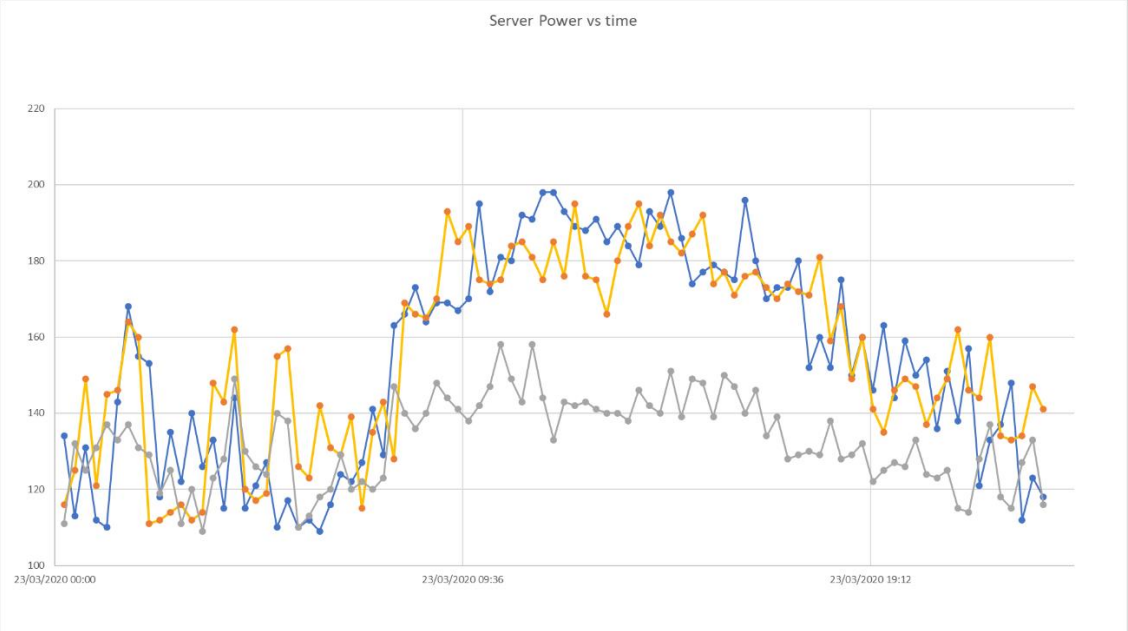


Figure 17: Measurements under high performance and balanced

Interestingly, the Idle power draw during the early hours of the morning was not significantly changed, it still hovers around 110W, active power is however lowered considerably. Total Energy was lowered with 14% over a 24-hour period.

Average power yields E_{total} = 3,18 kWh
Since CPU loading of the system is unchanged: E_{idle} = 2,64 kWh

Consequently, SIC = 83%

The fact that the SIC has risen (which is unexpected) is caused by the apparently wrong assumption that daytime power draw is entirely attributable to CPU load.
The result on power is very encouraging, interestingly, at this low CPU loading, the effect of power management manifests itself Not on the pure IDLE power but by switching back to the idle state more quickly and thus apparently acting on the active power instead.

Given the results obtained with the balanced mode, it was interesting to test if the server would exhibit even lower energy use when the power setting in the OS would be set to "power save". The "power save" mode lowers the maximum performance capability of the server by limiting over clocking (turbo modes). In this case, with low utilization, the effect of preventing over clocking will not pose any problems since maximum CPU performance is never needed. On the other hand, since the top turbo modes are rarely requested, the effect on power draw will also be small. The most interesting part of the measurement was the idle period (midnight to 7 AM), it was the expectation that the power save mode would allow deeper C-states and thus curb server power draw during idle.

After discussion with the participant in question, it was agreed that a fourth measurement run would be done on Monday September 20th, 2020, employing the power save setting on the entire Hyper-V cluster. The results are displayed together with the previous measurements in figure 18.

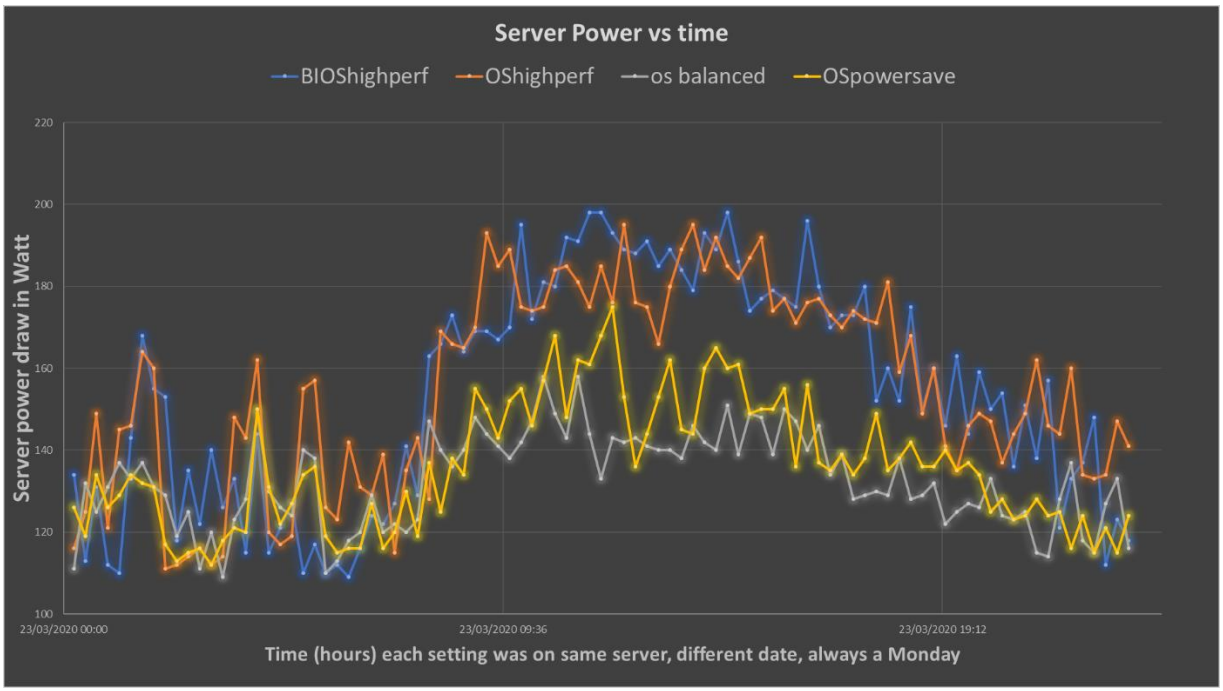


Figure 18: Measurements including "power save" mode

The results from the power save setting did not yield the result hoped for. Given the variations in power draw, the yellow line (power save) is not significantly different from the grey line (balanced). Specifically, during the midnight to 7 AM time period, the lines overlap. The small difference seen during daytime operations are most likely the result over small differences in CPU loading.

Again, no obvious impact on server performance was reported, however given the profiles it seems logical to suggest the balanced mode as the preferred setting for this server.

As a result of the measurements during the LEAP pilot, this participant has taken the advise and has permanently switched his HyperV cluster to the balanced mode.

4 QUALITATIVE ANALYSIS OF INTERVIEWS

After the analysis of the data, all 13 coalition partners participating in the pilots were contacted to discuss the analysis results, or when there was no data submitted, to discuss the reasons and barriers for providing data. Goal of these semi structured interviews was to ascertain the reasons for the choice of settings, if any changes in application performance had been observed during the second phase of the pilot, and what would be needed in order to broaden the application of power management.

What we observed during these conversations is that they are eager to improve energy efficiency of their ICT. The pilots often have raised awareness within these organizations of the energy consumption related to data handling. Having said that, delivering accurate data turned out to be more difficult than expected. It seems to be a combination of:

- Lack of technical knowledge to provide the correct pilot data:
  - It seems that there is a knowledge gap about the role of virtualization and power management in relation to costs and energy consumption. This leads to great inefficiencies in the setup / management of servers.
- Prejudices about power management:
  - All of the coalition partners cited the influence of power management on application performance as a reason not to change any settings. Even those parties that ran servers at a more stringent power management setting in week 2, reverted to the original settings without having observed a change in application performance.
- Lack of priority and policy:
  - Not many organizations have formulated policies around the use of power management, where these policies do exist, they most often state the use of high-performance modes.
  - We observed limited response despite instructions, available support (also from VMWare and hardware vendors) and reminders.
  - Hardly any use of the LEAP helpdesk - a total of 4 organizations contacted the helpdesk. Only a few organizations requested help from hardware and software vendors.

Hardware vendors are developing various technical or organizational solutions to increase the energy efficiency of hardware. Hardware vendors and software (operating system providers) are also aware that their training and instruction for the use of power management is not being used widely; it is available, but for many organizations difficult to find and understand.

5. WRAP UP

As part of LEAP track 1, a number of companies have supplied data that has proven useful for analyzing the potential of power savings by power management in servers. The data is however not conclusive. In certain cases, switching to dynamic power modes resulted in higher average power draw over a measurement period. In other cases, both OS controlled and hardware controlled power save mode resulted in clear energy savings.

Before presenting the observations however a note of warning must be provided. The number of physical servers that are currently active in the Netherlands is in the order of 1 million. The current study looked at 60 of these. Furthermore, these servers are not a random selection, they were selected by the LEAP coalition partners based on accessibility and/or specific non crucial functions that these servers have in the total IT infrastructure.

It is therefore not accurate nor realistic to quote averages above the level of single machines. The average CPU loading used in this document represent only the average of the associated machine and conclusions about the general state of ICT, the power settings and consequently savings potential cannot be drawn.

5.1 OBSERVATIONS

Not all of the data obtained in the pilot is discussed in the previous chapter, all data however was analyzed and is used as background for observations and conclusions.
The following table summarizes the data obtained, where appropriate, the reference to the corresponding paragraph is added.

IT-environment			OS		CPU%	Power [W]	Delta compared with measurement A	Comment
server type	workload	Bios-setting	layer	OS-setting	%	(avg)		
machine to machine	mixed	Static High Performance	Vmware	High Performance	39,3%	181		
machine to machine	mixed	OS-controlled	Vmware	Balanced performance	39,6%	205	13%	see paragraph 3.1
End user	mixed, average over all 20 nodes is shown	OS-controlled	Vmware	High Performance	18,0%	321		
End user	mixed, average over all 20 nodes is shown	OS-controlled		Balanced	17,0%	300	-7%	
machine to machine	monitoring	favor performance			41,0%	143		
machine to machine	monitoring	low power	Linux	throughput	41,0%	126	-13%	
machine to machine	monitoring	OS Control	Linux	power save	43,0%	160		
machine to machine	monitoring	OS control	Linux		43,0%	169	5%	there is not enough data to explain the discrepancy
enduser	VDI	high performance		n.a.	20,0%	190		a variation of machines were measured with varying settings for power management, the chosen example reflects dynamic high performance. Loading of this machine varried between 0 and 90% with a office hour pattern.
end user	frontend averaged 7 servers	Max performance	n.a.		5,4%	130		two servers show 1% average CPU load, severe underload
end user	backend averaged 4 servers	Max performance	n.a.		15,5%	580		average CPU load on one server 38%, other 3 below 10%
End user	3 servers on-site	Dynamic High Performance	Hyper-V	High performance	unknown, <<10%	154		
End user	3 servers on-site	OS controlled	Hyper-V	High performance	unknown, <<10%	155	0%	
End user	3 servers on-site	OS controlled	Hyper-V	Balanced performance	unknown, <<10%	132	-14,30%	savings when idle are 0%, office hour saving 23% saving
End user	3 servers on-site	OS controlled	Hyper-V	Powersave	unknown, <<10%	136	-12%	difference with balanced mode is negligable
enduser	storage	high performance	CentOS	n.a.	1,0%	72		storage server for HPC support (sporadic use)
enduser	storage	high performance	CentOS	n.a.	5,5%	228		storage server for HPC support (sporadic use)
enduser	mixed	high performance	CentOS	n.a.	9,0%	311		low utilization, storage and test vm's
enduser	HPC	high performance	CentOS	n.a.	79,0%	330		extremely high utilization consistent with HPC
enduser	HPC	high performance	CentOS	n.a.	79,0%	406		extremely high utilization consistent with HPC
enduser	mixed, averaged over 5 nodes	high performance	hyperV	n.a.	4,6%	346		extreme low utilization, customer is fixed on high performance due to advice from software vendor.
enduser	mixed averaged 5 nodes	high performance	Vmware	n.a.	40,0%	357		well utilized Vmware cluster, possibly profit from different power setting like case 2 cluster

Table 1 : data summary

The data obtained allows a number of observations that give usable insights into the usability of power management in servers and the savings that can be obtained in specific situations, both through the application of power management as well as from virtualization in combination with workload consolidation as has been described in previous reports:

- A majority of the respondent's report on servers that show dynamic power behavior. These modes result in a workload dependent power draw of these servers.
- All respondents apply some form of "high performance" setting by default, this has not changed due to the pilots.
  The reasons stated for applying the "high performance" are often semantic, the naming suggests in to be the most logical choice or advise from either hardware of software suppliers was to det the setting to high performance.
  Interestingly, this advice was often given many years ago. Sometimes in response to an IT incident sometimes as "a priori" advice. Since this advice was never retracted, the settings are inherited by current generations of servers.
- Many respondents apply conflicting settings on BIOS and OS level, coupled with the discussions held with the respondents, the root cause for these settings can be coupled to a lack of knowledge surrounding the settings for power management.
- Changing power management settings to more power saving modes results in approximately 10% energy savings in even highly occupied server nodes. No adverse effects on performance were reported during the testing of these power saving modes.
- Changing Static high-performance settings to dynamic high performance does not necessarily lead to energy savings.
- The best occupied servers still spend more than one third of their energy use on idle cycles, the worst occupied servers spend close to 99% of their energy on Idle.

Previous estimates on the possible effect of improved power management quoted higher possible savings than the observed 10% average, namely 20-40%. Firstly, in specific cases, such as the one reported on in paragraph 3.4, show that high savings can be obtained. In the specific case of the low utilization, the average power during daytime operations dropped by 23%, nighttime power was mostly unaffected. As such it is apparent that savings are workload dependent. What has also resulted in lower than previously estimated savings, is the fact that the "high performance" mode is no longer static. Almost all servers displayed dynamic power behavior even when set at "high performance". The savings from this dynamic behavior are already substantial, the observed 10% savings from the switch to a balanced mode can be seen as "additional".

It is also apparent (as a side-effect of the pilot) that virtualization has had a tremendous influence on power efficiency and still holds more potential than power management within servers. This is especially true for those servers that are (extremely) underutilized. Virtualization and consolidation of 10 or more of these servers can easily by done, resulting in not 10% but in an order of magnitude in power savings.

The highly utilized servers in our study supported up to 90 VM per physical node. If similarly, high levels of workload consolidation would be widespread, this would result in a very significant reduction in the number of active physical servers leading to massive reductions in energy use, let alone financial investments.

From the data we can also see that sustained high CPU utilization (up to 85%) does not result in problematic behavior. Several systems have continuous high loads and apparently the applications inside these machines run without faults.

Lastly, we have found that the Server Idle Coefficient (SIC) is a useful measure for denoting the energy waste in computing. It did not prove to be very sensitive to small changes in power management but is a promising measure for determining effective use of a server through both high CPU loading and very importantly, good dynamic ranges of these servers. The better the dynamic range, the better the server power draw follows the loading of the server. The best utilized servers had a SIC of 1,5 to 3, but the underutilized servers have SICs in the 20 to 100 range. Writing the SIC in percentages did not illicit much reaction, it is therefore more effective to write this in the same form as the PUE, namely as a number ranging from 1 to infinity.

5.2 CONCLUSIONS

Keeping the earlier mentioned caveats in mind, the following conclusions can be drawn on the basis of the pilot outcomes:

Further discussions with the respondents about the reasons and barriers for not applying power saving modes lead to very consistent answers: there are still major concerns about performance losses when applying power saving. Even if there is no evidence that this occurs.

In multiple cases, customer cite their software suppliers or system management parties that anything but “high performance” setting must be avoided.

Consistently deploying the “balanced” or “power save setting“ would contribute strongly to the goals for LEAP (i.e. improve energy efficiency of ICT in datacenters). Since none of the participating parties use a “power save” setting on their servers and a small minority used balanced, it is safe to assume that the use of these settings is uncommon in general. The use of the balanced setting yielded close to10% energy savings with two groups of highly utilized servers. It seems reasonable to assume that this 10% represents an estimate for the energy savings in all servers in use worldwide.

The potential for energy savings from consolidation of virtual environments is expected to be extensive. Pushing for higher levels can result in higher energy as well as financial savings than are currently targeted by the LEAP coalition. As shown by one participant, which deploys up to 90 virtual machines per physical node, with an average of 60 VM’s per node for the entire cluster, tremendous levels of workload/server consolidation can be achieved in a stable production environment. Applying similar levels of virtualized consolidation would decimate the number of servers needed to run the workload for most of the other participants.

The data collected implies further research is needed. The concerns about performance impact indicate better understanding and research into power management is needed, including impact on application performance. Also, a more comprehensive statistical analysis of the use of power management features and average CPU loading is needed to draw strong conclusion about the general use of power management features and the energy potential. Surprisingly, only half of the organization that initially committed supplied any data, again half of the data submitting parties did not change power management settings. When quizzed on this, it became apparent that lack of knowledge and fear of

consequences which is in itself again a result of a lack of knowledge prevented many organizations to participate in this pilot fully.

There is a pressing need for clear guidance and instruction from software and hardware providers, most ideally in unison, on how to best apply power management settings. This guidance must highlight the possible savings and explain when the standard power management can be tightened or must be relaxed.

Note that the use of “power management” and “virtualization” are measures within the framework of the “*Informatieplicht*” for data centers that are part of the “*Activiteitenbesluit*”. As stated in the observations, most parties deploy a dynamic power regime. Such a power regime can be seen as an application of power management.

5.3 RECOMMENDATIONS

The LEAP Track 1 pilots has yielded a wealth of information, both about the applicability of power management as well as on the human factor that controls the use of this feature.

From the combination of the raw data and the conversations with the participating parties a few recommendations for next steps emerge;

1) Guidance from hardware vendors and operating system providers.
2) Additional technical research into application performance under various power regimes.
3) Additional statistical research into the current use of power management and utilization rates.
4) Openness from (large) datacenters about the actual power draw of the facility over time.

The need for the supplier guidance has been discussed in the conclusions, the need for technical research must be construed as a supportive action. When data becomes available for the behavior of common applications running in various power regimes under carrying load conditions, it is likely that the currently ungrounded fear of a noticeable decrease of application performance will disappear.

Points 3 and 4 are also interlinked. There is currently too much uncertainty about the energy use of datacenters. The information is available, but not shared so conclusions based on total energy use as and information on time dependent power use cannot be drawn. Without this information as well as more detailed analysis on the actual use and configuration of ICT equipment, no accurate estimates can be made about the true potential of power management and virtualization that is still untapped within existing resources.

It would be useful to do a statistically relevant study into the actual power management settings as well as virtualization applied in servers in the Netherlands, from such a study more directed steps to unlock energy savings potentials can be created.