Power measurements
to increase power efficiency in data centers

Gyorgy Balazs

gyorgy.balazs@cern.ch

Supervisor:

Dr. Andreas Hirstius

andreas.hirstius@cern.ch

December 2008

Table of Contents

Summary

The document discusses the power consumption problems of data centers and possible solutions to improve their overall power efficiency. The main subject of the discussion is to show how power measurements can contribute in decreasing power consumption.

After a short introduction on the global problem of energy consumption in data centers, I present the Computer Centre of the European Organization for Nuclear Research (CERN), where power consumption became the largest challenge in the last few years.

In the next session, I give an overview on the different approaches to increase power efficiency in data centers, covering layout and cooling issues, hardware and even software solutions. After I have pointed out, why power measurements are essential when optimizing the power consumption, I show a practical example on how these power measurements can help increasing the power efficiency of data centers.

In the practical example, I describe my project to measure and analyze the power consumption of different computing server configurations, in order to help the procurement team of the CERN Computer Centre by calculating the power consumption of different hardware components, and identifying the most power saver configurations.

I show how I set up a test environment to measure power consumption, including planning, process and control development, validation. In the next chapters, I discuss my experience and problems during the measurements, and I give an overview on the gained results.

In the detailed analysis, I calculate the energy consumption of the main components inside the tested computers, using only the data obtained by measuring the power outside the servers on the main AC circuit.

In the last chapters of the example, I show my experience with several additional measurements, including blade and low-power solutions, compiler and platform optimization.

I conclude the document with an overview on my findings and a short note on the further development directions.

1. Introduction

The rising energy prices and our increasing hunger for electric power force us to take control over energy consumption. The demand for energy is rising heavily with the evolving industry and the more and more mechanized world, but the current primary energy sources are close to their limits, thus energy prices are going up very steeply. In all areas of industry a more and more crucial question became how to deal with the rising costs of energy.

The IT sector is also very sensitive. The energy prices are skyrocketing, but meanwhile the demand for computing power is also increasing. In some data centers the cost of electricity and cooling already exceeded the cost of the equipment itself. According to Gartner [8], power consumption and cooling issues became the single largest problem in 70 percent of the data centers. Gartner estimates that more than 50% of data centers will have insufficient power and cooling capacity to satisfy demands, while 48% of the overall budget is being spent on energy, up from 8% a few years ago. As Bob Worrall, chief information officer from Sun Microsystems says [11]: “There is no better time than right now to focus investment on more energy-efficient data centers.”

The purpose of this document is to give an overview on how data centers may significantly reduce their costs by taking measures to increase the effectiveness of power utilization. The aim of the document is not only to summarize the different techniques to increase power efficiency, but to show the importance of power measurements by looking at a practical example in detail. The example follows up a project that has been carried out in a remarkable data center to help increasing the power efficiency of installed servers by analyzing the power consumption of the different hardware components before the acquisition of new hardware.

2. The CERN Computer Centre

2.1 CERN – The European Organization for Nuclear Research

Power consumption became a key concern also at CERN in the last few years. CERN is the European Organization for Nuclear Research, the largest particle physics laboratory on the world, situated on the French-Swiss border near Geneva. CERN is the main site of high energy physics research in Europe, providing particle accelerators and other infrastructure for numerous research projects. The construction of the Large Hadron Collider (LHC), the new particle accelerator that will help to extend our knowledge on the fundamentals of the universe just have finished, and is going to start operation in 2009. The LHC with its 27km circumference, constructed 100m deep under the Swiss and French countryside is the largest particle accelerator ever built. The particle accelerator will regenerate the conditions that existed just after the Big Bang, by colliding particle beams in 4 detectors (ATLAS, CMS, ALICE, LHCb) along the ring of the LHC.

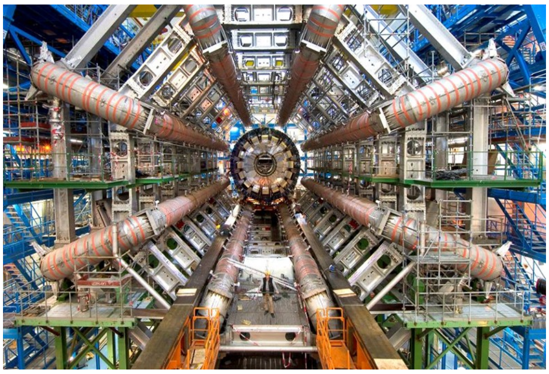


Figure 2.1
The construction of the ATLAS detector

2.2 Computing at CERN in the past and present days

These experiment sites produce approximately 12-15 petabytes data (equivalent to twenty million CD-ROMs) every year, which have to be processed and permanently stored. The Computer Centre is the core of data processing at CERN, equipped with state of the art computing facilities, storage and networking solutions, providing an enormous computing and storage capacity. Since the time when it was built in 1972, the Computer Centre has always been a pioneer in information technology. CERN was one of the first organizations introducing internet technology, and also CERN is the place where the World Wide Web was born to help scientists to share information based on hypertext documents. Recently CERN has become a centre for the development of Grid computing, hosting the Enabling Grids for E-sciencE (EGEE) and the LHC Computing Grid (LCG) projects. One of the two main Internet Exchange Points in Switzerland can also be found at CERN.

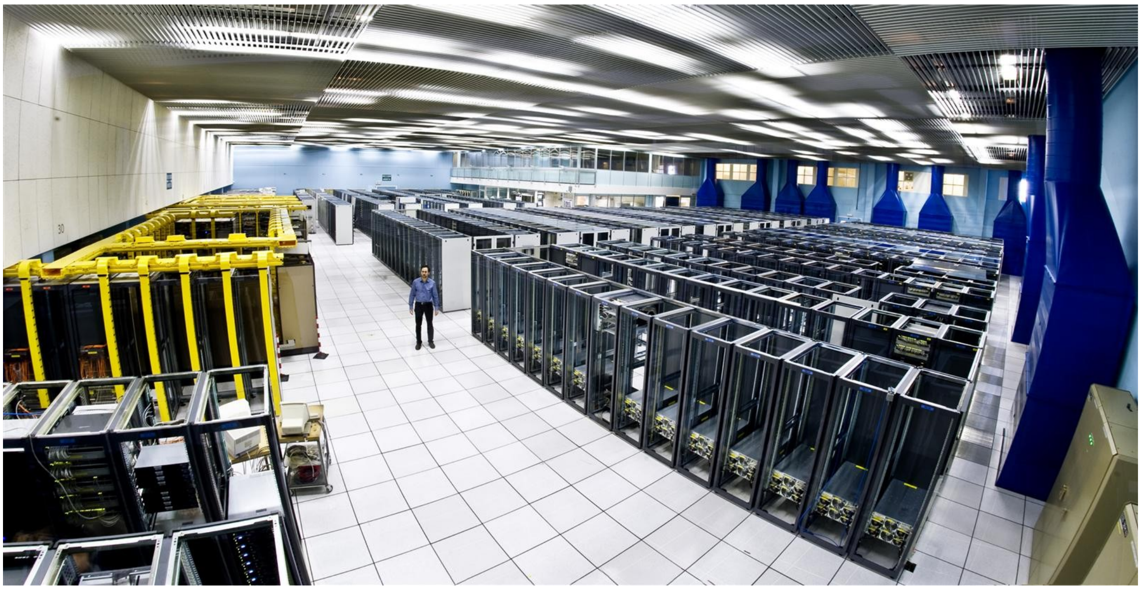


Figure 2.2 The CERN Computer Centre

The LHC experiment will produce an enormous amount of data, even tough the results will be filtered in real time on site, the filtered data still exploits the continuous 300 to 1200 MB/s real time transfer rate that is available for transferring the data to the CERN Computer Centre for storage and analysis. The data is cached on disk servers (about 4 petabytes capacity in early 2008) and the permanent storage is done on magnetic tape robots on site. Most of the data is analyzed in remote facilities all over the world with the help of the LHC Computing Grid, but a remarkable computing power is also available in the CERN Computer Centre. Currently more than 40000 processor cores are working in the PC farm, making up the core of the LHC grid, using cutting edge hardware and networking technologies.

2.3 The main challenge of the Computer Centre: Power consumption

One of the key challenges of the Computer Centre is the limited cooling capacity. The building was constructed to house large mainframe computers based on a design from the late sixties. But IT infrastructure has changed, and after refurbishing the Computer Centre to host high density rack mounted servers, the limited cooling capacity became the main detainer of further extension.. The building has a 2.5MW cooling limit, which currently translates into a 2.5 MW limit for the overall power consumption, which is expected to be fully utilized in the very near future. Therefore there is a very strong interest to optimize the power efficiency of the installed servers. Each Watt that is saved on the power consumption gives an additional saving on the cooling. Since the demand for computing power is generally infinite in the high energy physics community, all power savings are used to give place for new equipment.

3. Strategies to increase the power efficiency in data centers

Since power consumption and cooling issues became a key problem in computer centers all over the world, various measures have been introduced to address the problem. Several techniques can increase power efficiency including layout modifications, cooling approaches, hardware and even software techniques. Based on the experience at CERN [1] supplemented with the recently developed techniques, there are plenty of approaches to follow:

3.1 Thermal issues and data center layout

The design of the data center layout and cooling has a major influence on power efficiency. Blowing cold air from the ceiling was a sufficient approach in the time of large mainframe computers when a moderate temperature in the room was enough to keep the single computer cool. But the density of computers have changed, and now extreme heat must be extracted from such small areas as a fingertip. There are several approaches to provide efficient and precisely controlled cooling solutions. The most important techniques:

- Blow the cold air from the raised floor close to the racks instead of cooling from the ceiling.
- Avoid hot and cold air mixing by creating hot and cold aisles. In a cold aisle, the racks are aligned to face each other, and cold air is blown in between them, while the hot air is extracted to the hot aisle at their back.
- Seal the cold aisles, so the cold air is kept inside instead of blown up above the racks.
- Extract warm air from hot aisles
- Place cooling as close to the equipment as possible. Row or rack oriented cooling solutions provide high effectiveness. Using racks equipped with water cooled heat exchangers can increase power efficiency and provide considerable solution even in environments with high density equipment and limited air cooling capabilities.

- The large number of small fans that are spinning very fast in 1U size rack mounted computers tend to be much less efficient than using a smaller number of large fans in enclosures with a larger form factor as 2U, 4U. Especially blade solutions provide power efficient air cooling inside the enclosure with large, collective cooling fans.
- Looking at the long term trends, new in-server cooling solutions will also emerge. Air cooling is not sufficient anymore as the density of computers increases, but providing proper traditional liquid cooling is also challenging. There is already an alternative solution from SprayCool Inc. [6] that extracts the heat from any hot surface by using the latent heat of evaporation of a liquid.

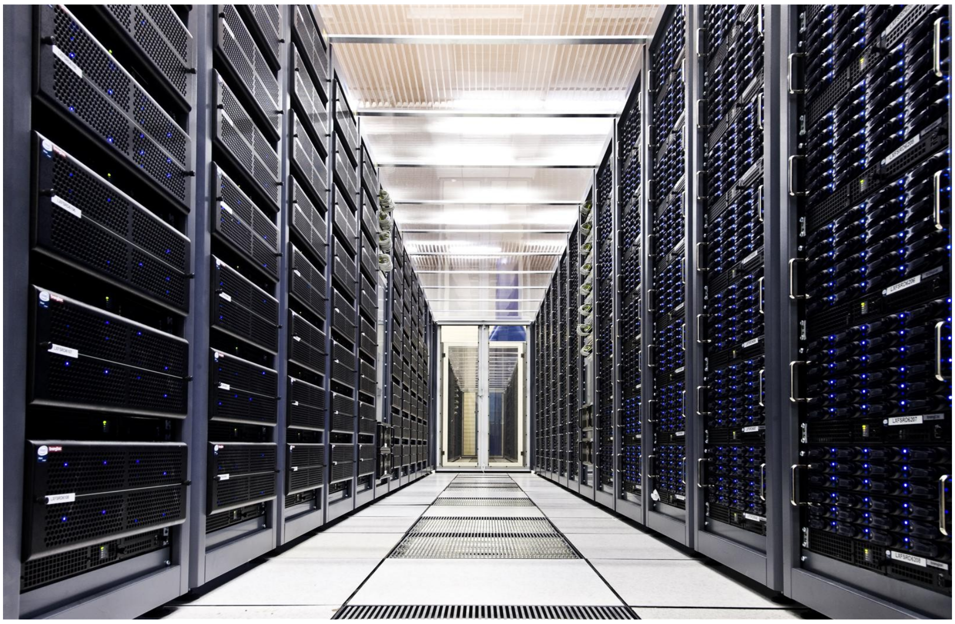


Figure 3.1 Cold aisle with closed sealing in the CERN Computer Centre

3.2 Power supplies

The overall power efficiency of a computing system is highly influenced by the way it provides power to its components. The incoming AC voltage is transformed to DC voltage by a given loss of energy. According to our measurements, a standard desktop power supply provides as low efficiency as 50-70%, while the high quality power supplies for server computers provide efficient power conversion with a ratio up to 99%. In general, larger power supplies tend to be more efficient. In a blade system, a smaller number of larger power supplies provide redundant power source to all blades in the enclosure with only a minimal loss of energy.

Redundancy is essential in a server environment, but introducing more power supply redundancy than it is necessary also comes with less efficient power conversion. By sharing workload, each power supply has to provide less energy, but in most case that reduces also their effectiveness.

Intelligent power supply management is also being developed. Equipping power supplies with digital controllers allows to collect data about the current state of the supplies (heat, operating time, load response), allows dynamic load balancing and up-down sequencing. [3]

Another approach is to provide central AC to DC conversion for the whole data center, so only DC to DC conversion is needed for the computing nodes. Transforming the power centrally improves the efficiency very much, but also has several drawbacks. Besides the very challenging practical realization of such a system, providing scalability is also a difficult task in this case.

3.3 Hardware configuration

Choosing the appropriate hardware is a key issue when increasing data center power efficiency. Currently each Watt saved on the hardware configuration translates into an additional Watt saved on the cooling. There are two main energy consumers in a computer that are also good reserves for power savings. The main consumers are traditionally the CPUs that are dissipating 50 to 150 Watts depending on the architecture, but since a high level of parallelization has been introduced in computing, with the increasing number of parallel processes, also the hunger for memory has multiplied. At many organizations such as CERN, adding more execution units to the system means proportional extension also to the memory. Thus the power consumption of the main memory became not only comparable to that of the processor, but in some cases even overpasses it. The later described investigation addresses the question of choosing power effective processors and memory configuration in order to deploy more power efficient server computers.

3.4 Software techniques

- Virtualization

  It is possible to run more than one operating system as logically separated entities on the same hardware by using a virtualization software. Most times each physical server represents a service, but this may lead to underutilization of the underlying hardware, since many services are not used constantly. Merging these logical services into a virtualized environment using less physical hardware may result in large power savings.

- Platform optimization

  Today's hardware solutions provide several opportunities to optimize performance. The frequency of the CPU can be adjusted according to the actual workload by native or external software or IPMI device, thus saving power when the system is idle. The Intel EIST (Enhanced Intel Speedstep Technology) provides dynamic frequency and core voltage scaling that can be activated from the BIOS. Using the different power saver 'sleep' modes in idle systems (e.g. switching off hard drives) may result in noticeable power savings.

- Multi-threading

  Multi-threading is a software method that earns more importance since processors with more and more execution slots (multicore, hardware-threaded CPU-s) are evolving. Multi-threaded applications may lead to more efficient utilization of the hardware, thus significant savings on energy.

- Compiler technologies

  The compilers that are creating binaries from source code often influence the performance of the generated executable. There are optimizing compilers, such as the Intel C/C++ compiler (icc), that are able to optimize the code for the underlying hardware platform reaching significant speed up, resulting in better power efficiency.

4. Power measurements

To be able to handle energy costs and increase power efficiency, the first step to be made is measuring the power that is currently consumed by all the facilities. Understanding the current environment is essential to be able to assess risk factors and take actions to increase efficiency. While even a small increase in efficiency can make a a difference in energy costs, a focused plan of action can result in significant savings.

Collecting accurate, aggregated energy consumption data is a challenging task in a data center, since hardware is always in change, and also workload fluctuations introduce significant changes in the measured values. Using solid metrics is also important when measuring efficiency.

4.1 Power efficiency metrics: PUE and DciE

There are several metrics to define overall data center power efficiency, but recently the metrics proposed by the Green Grid organization has become widely accepted. These are the Power Usage Effectiveness (PUE) and Data center Infrastructure Efficiency (DciE) [2]

The PUE is defined as follows:

PUE = Total Facility Power /IT Equipment Power

and its reciprocal, the DCiE is defined as:

DCiE = 1/PUE = IT Equipment Power x 100% /Total Facility Power

The components for the loads in the metrics are described as follows:

1. The IT Equipment Power is defined as the equipment that is used to manage, process, store, or route data within the data center. This includes the load associated with all the IT equipment, such as computers, storage, and network.

2. The Total Facility Power is defined as the power measured at the utility meter — the

power dedicated to the entire data center, including IT equipment and the supporting devices:

- Power delivery components (UPS, generators, PDUs, batteries, distribution losses)
- Cooling system components (chillers, room air conditioning units (CRACs), etc.)
- IT equipment (Compute, network, and storage nodes)
- Other miscellaneous component loads such as data center lighting

The average power efficiency of data centers is 0.3, indicating that only 30% of the consumed power used by the IT equipment. The following figure shows the distribution of the consumed energy in a usual data center.

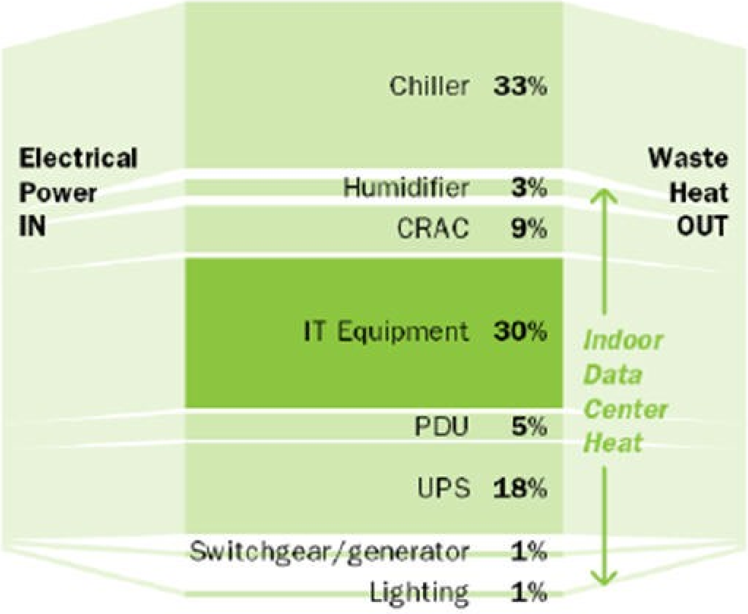


Figure 4.1

Distribution of electric power in average data centers

The PUE and DCiE provides a way to determine:

- Opportunities to improve a data center's operational efficiency.
- How a data center compares with competitive data centers.
- If the data center operators are improving the designs and processes over time.
- Opportunities to repurpose energy for additional IT equipment.

The power consumption can be measured on many levels of the equipment hierarchy starting from the utility power meter of the data center, down to the power consumption of individual computers. Measuring power on lower levels gives more accurate results and allows for detailed modeling, but also increases the time and costs spent on the measurements. CERN performs low level power measurements on the equipment, measuring the individual servers not only after installation, but also before the acquisition of the new hardware.

4.2 Power measurements before acquisition as a measure to save power

CERN already included the power consumption into the TCO (Total Cost of Ownership) calculations of server computers, and encourages the distributors to provide power efficient solutions for the tendering projects. During the acquisition process, an evaluation system which represents the final configuration to be offered is tested at CERN, and among performance tests, also power measurements are being performed. The consumed power is measured under both idle and full load. After the measurements a value of 80% load and 20% idle result is generated, which represents the average power consumption of each evaluation system.

Before the price comparison of the offered systems, a financial penalty is added to the original price, which represents an estimation of all electricity costs for a 3 years lifetime. This financial penalty currently amounts to 10 CHF (Swiss francs) after each Watt of the average power consumption. This way, 5000 CHF penalty is added to the price of a system with 500W average power consumption, and the final decision is made according to the calculated final prices. Including the electricity costs in the evaluation process strongly motivates the distributors to provide the most power efficient solutions available.

5. A project to increase power efficiency by analyzing the power consumption of main hardware components

Recognizing the practical value of power measurements, a project has been started to measure and analyze the power consumption of a broad selection of probable hardware configurations. The aim of the project is to gain a better understanding on server computer's power consumption and help procurement to increase power efficiency by identifying the most efficient solutions during the acquisitions.

One very important aspect for increasing the power efficiency is to gain a detailed knowledge about the influence of different components of the servers on power consumption. The focus of this investigation is on the two major power consumers in a server, the CPUs and the memory. Traditionally the CPU has been the main energy consumer, but with the new generation of very efficient multi-core processors, the power consumption of the memory comes into focus now.

During the measurements, different CPU and memory setups are examined on the same base systems. The power consumption of the test systems is measured directly in the main AC circuit of the machines while different benchmarking programs are running to generate several different load conditions on the systems. Each test program stresses the hardware in a different way. using different resources. The power consumption of the machines changes under each test, allowing for a detailed analysis of each hardware component.

The results of the measurements with different CPUs and memory modules provide enough information to analyze the individual power consumption of the CPU and the memory. The values for a particular CPU or memory module can be derived from the comparison of the results of all corresponding configurations under the different load states. The goal of all calculations is to give an accurate estimation for the power consumption of each type of the tested memory modules and CPUs. This information gives us better understanding on energy usage and it will help us building more power efficient systems in the future.

5.1 Test environment setup

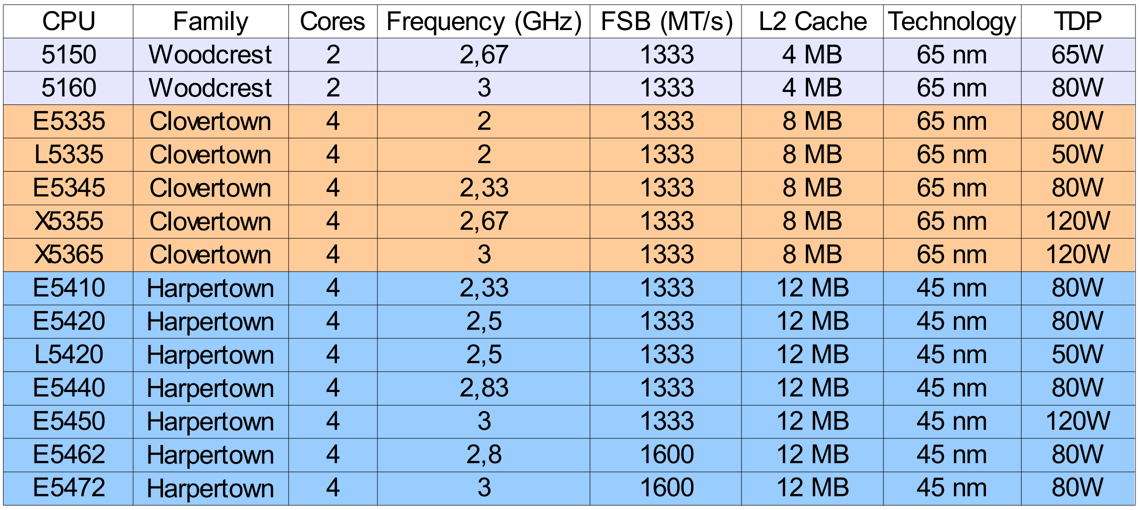
The tests are to be conducted in a test environment, where the results are not influenced by the outer world neither by any anomaly that can occur in a single system. To ensure that, the tests are performed in a lab environment with two, initially identical test systems.

The test systems are to be initially installed with completely identical hardware configuration, having the same model of motherboard, CPUs, memory, power supply and chassis. A very important requirement, that after the installation, the identical configurations should consume the same amount of power. If it is proven, that the systems are identical also in power consumption, the test process can be sped up by performing parallel measurements with different CPU and memory configurations.

5.1.1 Hardware configuration

The test covers all the dual and quadcore CPUs and FB-DIMM (Fully Buffered DIMM) memory modules on the market which are of interest to CERN. In order to conduct all the tests in the same environment, the base system should support all current Core2 based CPUs (Woodcrest, Clovertown and Harpertown) and all the memory modules that are being investigated. For that purpose the dual-socket SuperMicro X7DWN+ motherboard equipped with Intel's Seaburg chipset and 16x FB-DIMM slots have been chosen.

CPUs that are taking part in the measurements:



CPU	Family	Cores	Frequency (GHz)	FSB (MT/s)	L2 Cache	Technology	TDP
5150	Woodcrest	2	2,67	1333	4 MB	65 nm	65W
5160	Woodcrest	2	3	1333	4 MB	65 nm	80W
E5335	Clovertown	4	2	1333	8 MB	65 nm	80W
L5335	Clovertown	4	2	1333	8 MB	65 nm	50W
E5345	Clovertown	4	2,33	1333	8 MB	65 nm	80W
X5355	Clovertown	4	2,67	1333	8 MB	65 nm	120W
X5365	Clovertown	4	3	1333	8 MB	65 nm	120W
E5410	Harpertown	4	2,33	1333	12 MB	45 nm	80W
E5420	Harpertown	4	2,5	1333	12 MB	45 nm	80W
L5420	Harpertown	4	2,5	1333	12 MB	45 nm	50W
E5440	Harpertown	4	2,83	1333	12 MB	45 nm	80W
E5450	Harpertown	4	3	1333	12 MB	45 nm	120W
E5462	Harpertown	4	2,8	1600	12 MB	45 nm	80W
E5472	Harpertown	4	3	1600	12 MB	45 nm	80W

Figure 5.1 CPU types to test

Most of the processors to test are standard mainstream CPUs marked with (E)5xxx. The so called “Extreme” processors (X53xx) which provide extra computing power by reaching higher frequencies but also have a higher Thermal Design Power (TDP) are tested as well. There are also processors available with a lower TDP than standard CPUs, the “L” series of Intel Xeon CPUs. From the “L” type CPUs the L5335 and L5420 are to be tested.

Memory configuration:

The amount of memory used in the servers is determined by the number of cores in the CPU. The current requirement from the LHC experiments is 2GB per Core, so for each CPU core in the system 2GB memory needs to be added. That way a total amount of 16GB memory is used with the quad core Clovertown and Harpertown systems. (2 CPU * 4 cores * 2GB memory)

The tests with the 2 dualcore CPUs are performed both with 8GB and 16GB memory, but for the tendering process, the results with 8GB memory (2 CPU * 2Cores * 2GB memory) will be used.

Notations that are used to mark the actual memory configuration in the tests:

667MHz FB-DIMM modules
1G@667 1GB 667MHz FB-DIMM modules – 16 Modules per system (8 for 8GB)
2G@667 2GB 667MHz FB-DIMM modules – 8 Modules per system (4 for 8GB)
4G@667 4GB 667MHz FB-DIMM modules – 4 Modules per system (2 for 8GB)

800MHz FB DIMM modules (With 1333MHz FSB CPUs running on 667MHz)
1G@800 1GB 800MHz FB-DIMM modules – 16 Modules per system (8 for 8GB)
2G@800 2GB 800MHz FB-DIMM modules – 8 Modules per system (4 for 8GB)
4G@800 4GB 800MHz FB-DIMM modules – 4 Modules per system (2 for 8GB)

5.1.2 Software configuration

The test systems are installed with the Red Hat Enterprise based Scientific Linux CERN 4.6 with 2.6.25.1 kernel. The installation is performed with the default options and packages for any production batch system at CERN. Only the test programs for the measurements are installed additionally.

5.1.3 Power meter

The measurements are performed with a *ZES Zimmer LMG 500* power meter. The power meter is controlled by a laptop that is connected via the RS232 interface. The test systems obtain power through the measurement adapter (LMG-MAK1) which enables the power meter to perform accurate measurements directly in the main ~220V AC circuit. The laptop samples the measured values in every 10 seconds and stores them in a simple comma separated text (csv) file. The whole process is controlled centrally by scripts on the remote workstation which uploads and starts the test programs on the measured systems and controls the power meter at the same time.

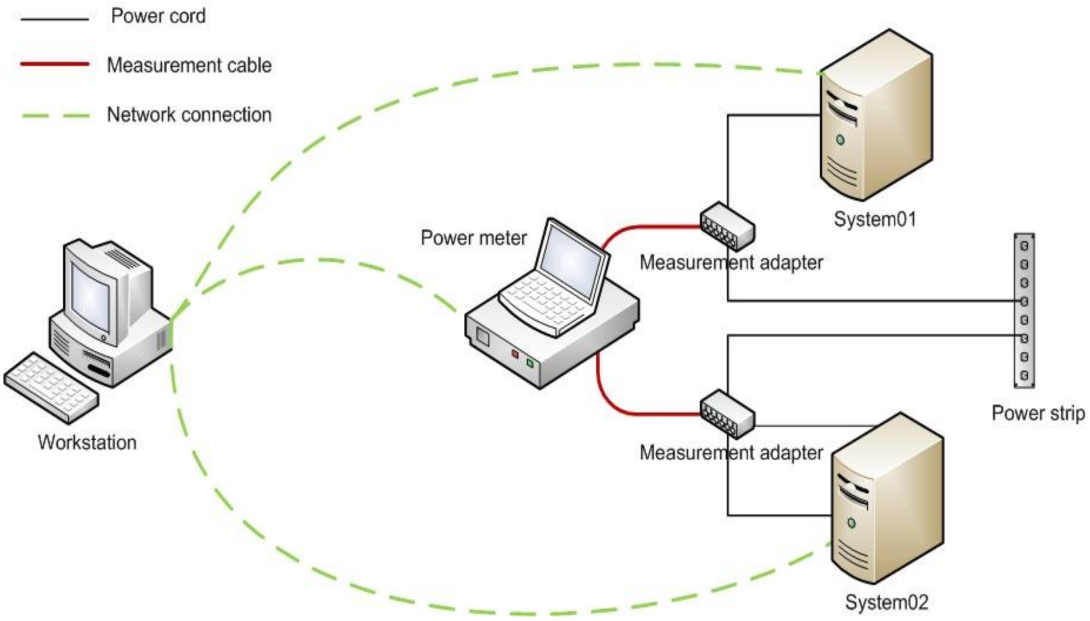


Figure 5.2 The layout of the measurements

5.1.4 Measured values

The following values are measured during the tests [9]:

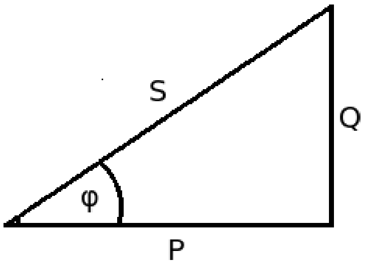


Figure 5.3

Components of electric power

- Active Power (P): The component of electric power that performs work, often referred as 'real' power. That part of the electric power can be used by the actual power consumers. It is measured in Watts (W).

- Apparent Power (S): The product of the voltage (in volts) and the current (in amperes). This part of the power represents what is being drawn from the electrical circuit, it comprises both active (P) and reactive (Q) power. It is measured in volt-amperes (VA).

- Power Factor: The ratio between the Active power and the Apparent power. The power factor is a number between 0 and 1 representing the power efficiency of the power supply in our case.

Since the overall electric power consumption is being invoiced regarding the Apparent Power that is measured at the end of the subscribers power line, the results for the Apparent power consumption is used in CERNs tendering process. For deeper analysis and calculations to examine the power consumption of the different devices in the computers, the Active Power results are used.

5.1.5 Test programs

The following test programs are used to stress the hardware while the power measurements are being taken:

- CPUburn: is designed to load the CPU as heavily as possible. It is part of the standard CERN test toolkit for power measurements during the tendering process for new servers.
- Lapack: is designed to load the memory subsystem and also generate load on the CPU. It solves a very large linear equation. Lapack is also part of the standard measurement process in tenders.
- The HLT Test program: The test is derived from the High Level Trigger (HLT) that is used at the ALICE detector of the LHC experiment. The original program filters the useful information from the data generated by the detector when collisions occur. The test program generates utilization on the test system which is close to the actual utilization of machines in full production, so the power consumption for production systems can be estimated.

The HLT benchmark is available in different versions.

Single threaded versions:

- single – 32 bit SSE instructions: single precision, single thread
- double – 64 bit SSE instructions: double precision, single thread
- x87f – 32 bit x87 instructions: single precision, single thread
- x87d – 64 bit x87 instructions: double precision, single thread

Multithreaded versions

- tbb1 – 32 bit SSE instructions: single precision, multithreaded – 1 thread
- tbb #c/2 - 32 bit SSE instructions: single precision, multithreaded – number of cores/2 thread
- tbb #c - 32 bit SSE instructions: single precision, multithreaded – number of cores thread

Running the program in the different versions allows us to test for example different execution units in the processor.

5.1.6 Test process composition

All possible CPU and Memory combinations are examined using the following set of tests. The goal is to set up a test process that generates workloads of different type and intensity on the tested system. The tests should always take the same amount of time and follow each other in the same order.

The complete test process:

- Idle test:

  60m **idle**

- Load test: (#c = number of cores)

30m **mixed cpuburn+lapack** #c times

15m idle

30m **cpuburn** #c times

15m idle

30m **single** #c times

15m idle

30m **double** #c times

15m idle

30m **x87d** #c times

15m idle

30m **x87f** #c times

15m idle

30m **tbb 1** times 1 thread

15m idle

30m **tbb #c/2** 1 times #c/2 thread

15m idle

30m **tbb #c** 1 times #c thread

The tests, except the multithreaded HLT tests (tbb), are executed as individual processes on each core of the system: 8 times on the systems with 2 quad core CPUs and 4 times on the dual core systems.

The 3 most important tests are the idle, cpuburn and the mixed cpuburn+lapack tests. The detailed analysis of components is based on the power measurement results of these tests.

Idle test: The power consumption is measured while only the standard operating system processes are running, all components are considered to be in idle state.

Cpuburn: Generates an artificial full load only on the processors while the memory subsystem remains idle. That way, the power consumption of the CPUs can be derived from the results (taking into account that the base system has also a higher load).

Mixed cpuburn+lapack test: Runs cpuburn and lapack on alternating cores by pinning manually each process to a physical core so, that the load on both sockets remains balanced. The test generates full load on both the CPUs and the memory subsystem, so the power consumption of the memory can also be observed.

The following plot shows the active and apparent power results of a generic load test:

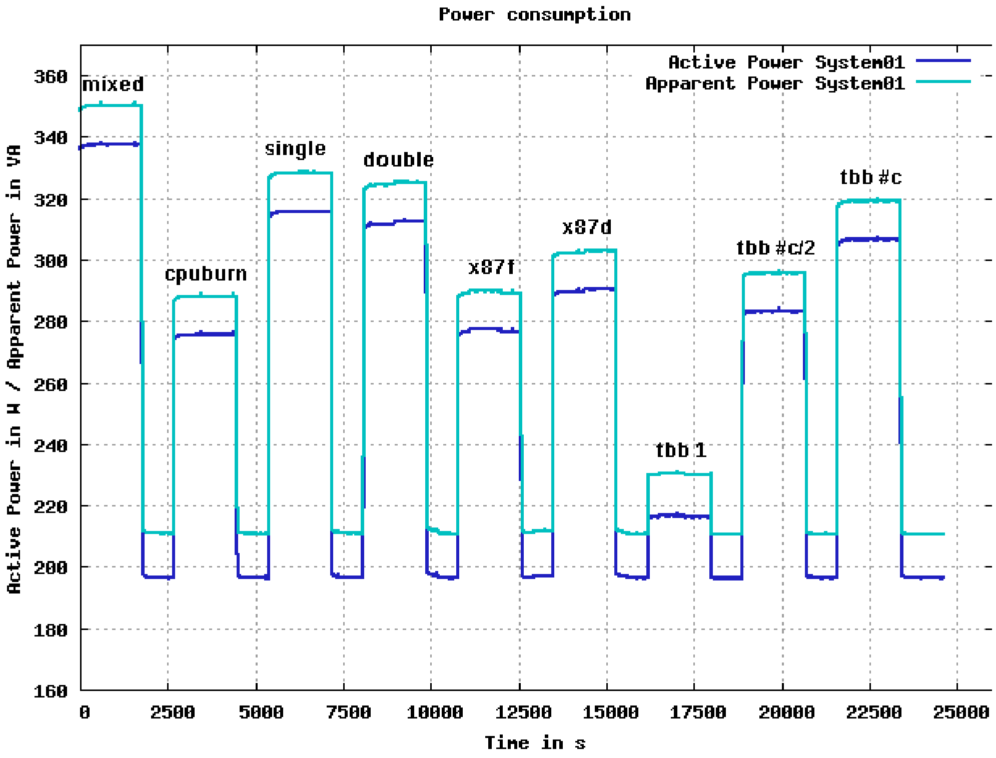


Figure 5.4 Generic load test results, Active and Apparent power

5.1.7 Automating the test process

A very important goal is to assure that the complete test procedure is executed on the same way on all tested configurations. To achieve that, the whole test process needs to be automated. The solution is to script all tasks to be done, so the measurements can be controlled centrally from a remote workstation.

Tasks to be done:

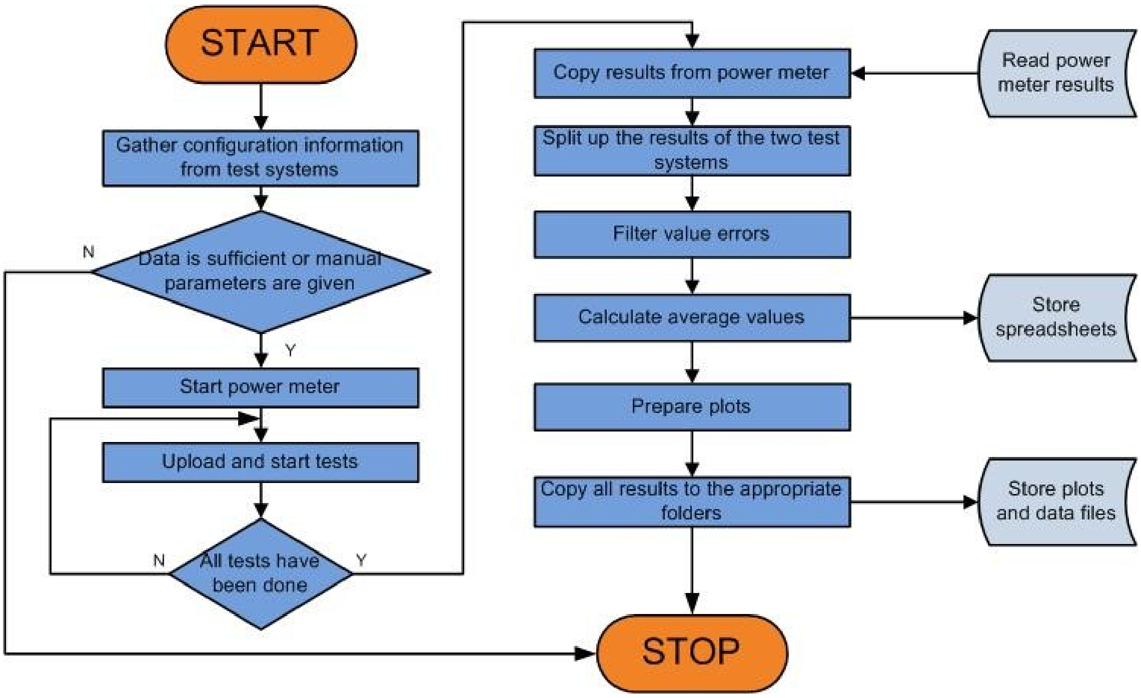

START
Gather configuration information from test systems
Data is sufficient or manual parameters are given
N
Y
Start power meter
Upload and start tests
All tests have been done
N
Y
Copy results from power meter
Read power meter results
Split up the results of the two test systems
Filter value errors
Calculate average values
Store spreadsheets
Prepare plots
Copy all results to the appropriate folders
Store plots and data files
STOP

Figure 5.5 The measurement process

5.1.8 Control process development

- Startgui.py – It starts a graphical user interface, where the actual processor and memory setup can be selected for both machines from a list. The lists are filled up from the cpu.txt and memory.txt files which contain all the possible cpus and memory modules. It is also possible, to start the tests with special parameters by filling an additional tag field. The program calls the exec.sh script with the given parameters.

- Exec.sh – The script controls all the test process from performing the measurements to filling up spreadsheets with the averages of the gained data. Each relevant task is done by a different script to increase reusability.

  Tasks to be controlled by the exec.sh script:

  - Gathering actual CPU and memory setup information from the machines, which will be stored in memconf.txt and cpuinfo.txt
  - Call standard_test.sh – runs the tests on all the measured systems in parallel while controlling also the power meter. It performs 1h idle test, 7h load test and the results are generated on the power meter laptop.
  - Copy results from the power meter to the “processing” temporal directory
  - Call csvprocess.py – Since the consumption values for all the machines that are tested simultaneously are stored in a single CSV (Comma Separated Values) file, it is needed to split up the results to store the data in a different file for each machine. csvprocess.py splits up the results into two CSV files, and filters all false values avoiding errors made by the power meter.
  - Call spreadsheet.py – Reads through the results, and calculates averages for each and every test that has been made. After detecting a rising edge, it starts averaging after the 10th minute until the next falling edge. The last 10 seconds are also cut down from the average numbers. It saves the results only if the number of high edges are equal to the number of tests. The results are stored in three different location:
    - cpusheets: There is one spreadsheet for each processor, and each row contains the results for a different memory configuration
    - testsheets: There is one spreadsheet for each test type, and each row contains the data for the corresponding test of a different cpu/memory configuration
    - allresults: All results are also stored in a 'universal spreadsheet', where each row defined by the cpu, memory and the testname property.

- Makeplots.sh – Creates plots out of the corrected raw data. For each machine and each data file the following plots are created:
  - ■ Active and Apparent power
  - ■ Power factor
  - ■ A few comprehensive plots are created from the data of both machines to allow a fast review and error detection.
- Copy all data files to the appropriate place.
  - ■ Raw data files are stored in power/results directory
  - ■ Plots are stored under power/plots directory
  - ■ Spreadsheets are stored in the sheets directory

An overview of the process:

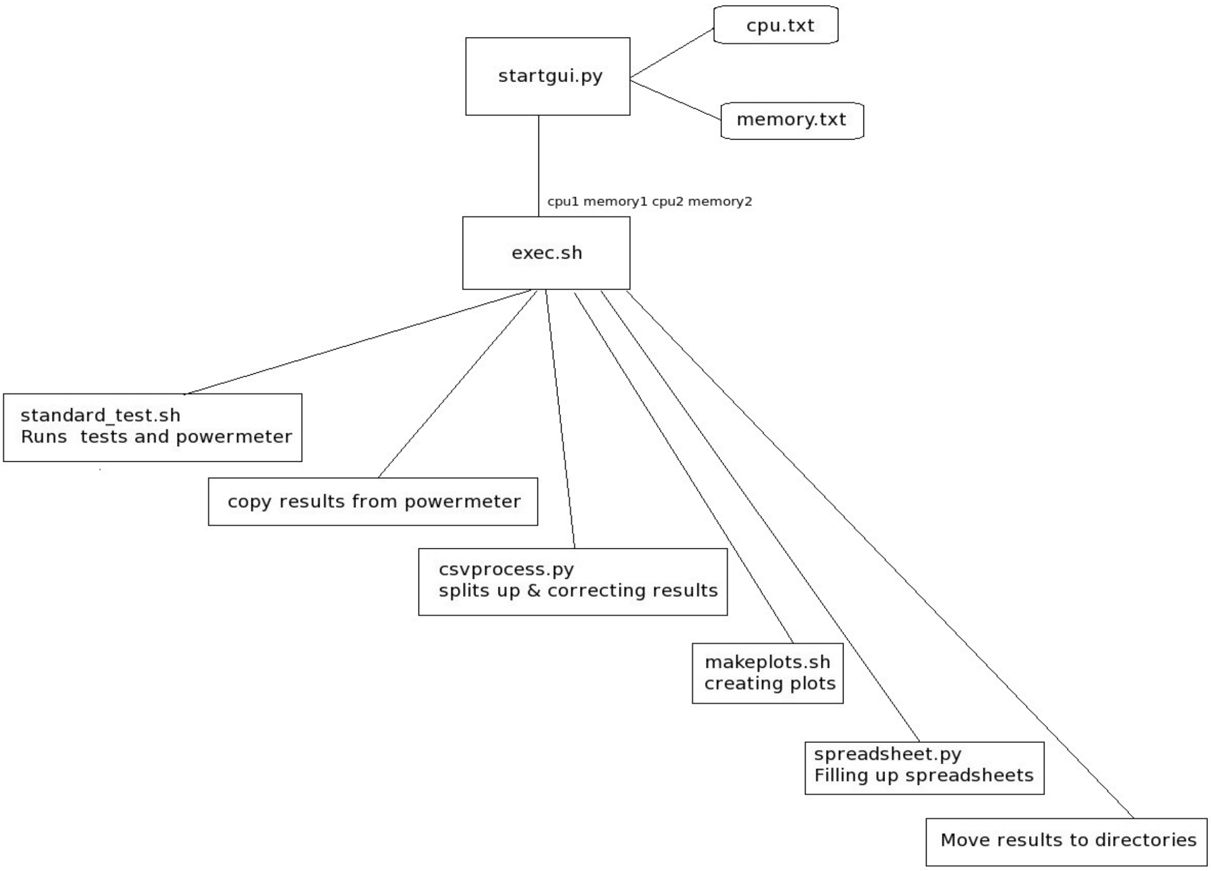

cpu.txt
startgui.py
memory.txt
cpu1 memory1 cpu2 memory2
exec.sh
standard_test.sh
Runs tests and powermeter
copy results from powermeter
csvprocess.py
splits up & correcting results
makeplots.sh
creating plots
spreadsheet.py
Filling up spreadsheets
Move results to directories

Figure 5.6 Control process overview

5.1.9 Validating the test environment

It was a very important requirement to set up a test environment where both systems are identical in their hardware and software configuration and work under the same environmental conditions.

The two test systems are required to have identical:

- Hardware configuration
- BIOS setup
- Operating system
- Software installation
- Temperature
- Cooling conditions

The test systems were purchased with identical hardware configurations and after the installation that was done according to the above declared requirements, several tests were performed to examine if they are also identical in power consumption.

The first results showed about 1% difference in power consumption both in idle and load state, and a noticeable difference appeared between the machines during the mixed cpuburn+lapack test. This discrepancy was caused by a failure in the manual pinning of the jobs to cores. The problem has been solved. Detailed description of the problem and the solution can be found in the "Problems during the measurements" section of the document.

After the systems had stabilized, the tests were repeated several times, with identical results. Thus it was proven that the machines consume the same amount of power under the same circumstances. This allows to perform parallel tests with different hardware configurations, cutting the required testing time in half with an estimated maximum error of +/- 2 Watts for the whole measurement process.

The following plot shows the aligned Active power results for both test systems with identical configuration after all stability problems had been solved.

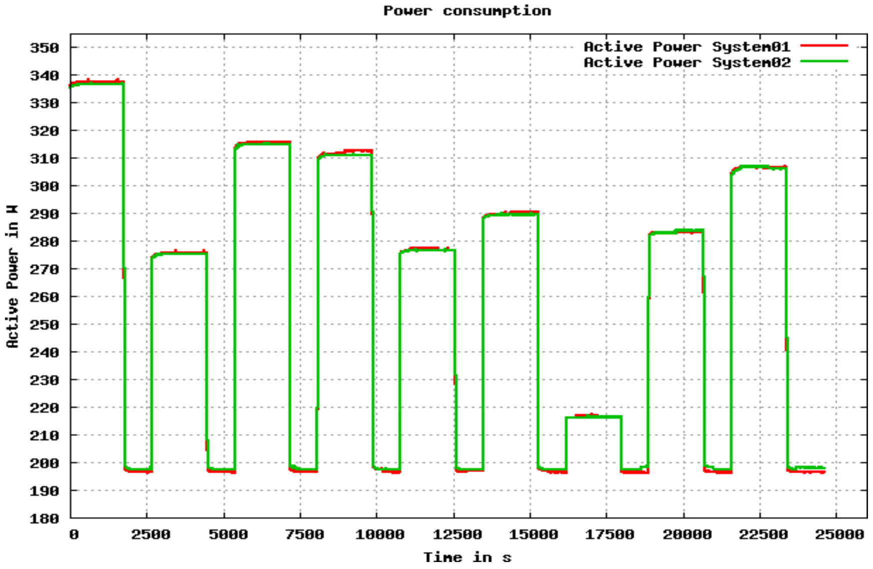


Figure 5.7 Generic load test after final installation, Active power

As the plot shows, the results are completely identical under all load conditions.

After all measurements were done, another test was performed to recheck the difference between the machines. The test was successful, the results showed no change compared to the initial measurements.

5.2 Performing measurements

The current investigation includes 14 processors and 3 different memory module sizes. All possible configurations were to be tested., and additionally several measurements have been done also with the higher frequency (800MHz) memory modules.

Each configuration went through the 8 hours long test process that is described above. The measurements were performed during 2 months, by starting in most cases one process for the day, and one for overnight. A shorter test process would have been enough to compare the results and also for the basic calculations, but in the current investigation it was very important to collect as much data as possible, also to provide data for further analysis.

5.3 Problems during the measurements

5.3.1 Manual process pinning failure

The first measurements were conducted using two identical configurations to set up a reliable testing environment. The results showed a noticeable difference between the two test machines under the mixed cpuburn+lapack test, and the several times repeated test showed, that there are also differences on the same machine between the different runs of the test. The plot shows the aligned results of several test runs.

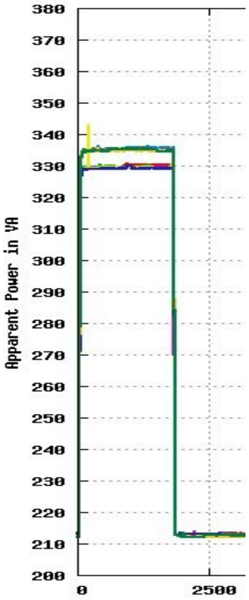


Figure 5.8 Aligned results, Apparent power

Reason

After examining the script code of the test, it was found, that there is a failure in the manual pinning of the jobs to the different CPU cores. In worst case it was possible, that one physical CPU was running all the cpuburn jobs meanwhile the other CPU was running only lapack jobs.

The failure was caused by the false bitmask in the taskset command which is used to perform the manual pinning of processes to the bitmask specified CPU core.

Solution

Instead of using the bitmask, the script has been changed so, that the taskset command is executed with the -c #coreid parameter using a numerical identifier to specify the corresponding core for the actual process. That way the cpuburn and lapack jobs are executed on alternate cores and the load between the CPU sockets are also balanced.

After changing the manual pinning process in the script, several tests were performed to prove if the results are identical on each machine, and the difference between the machines have been also stabilized.

5.3.2 Unexpected activity on IPMI device

The output of the idle power measurement on one of the test systems showed unusual jumps in power consumption. The power measurement was taken in idle status, which means that only standard operating system processes were running on the system. The unexpected jumps were repeated in every 2 minutes during the whole test process.

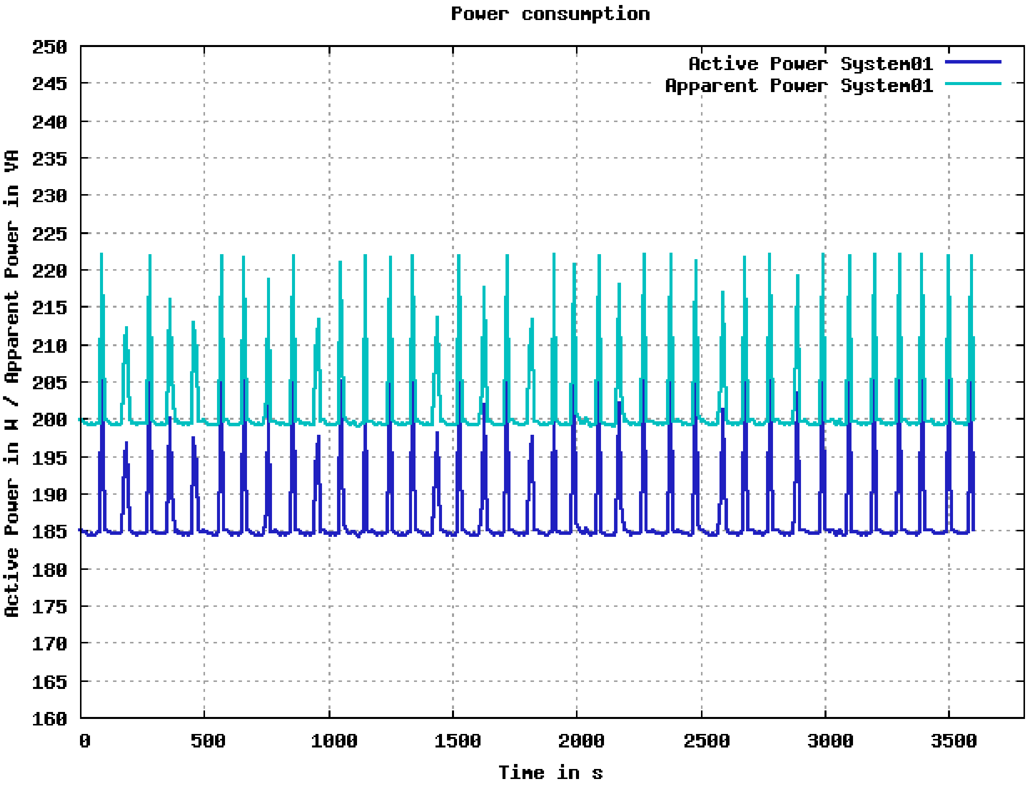


Figure 5.9 Unexpected activity under idle measurement

<u>Facts</u>

- Checking running processes with 'ps -ax': Only the standard OS processes were running
- Real time monitoring of processes with 'top': Only standard OS activity could be seen, except that 'udev' appeared for some seconds in every two minutes

- In the system log the following messages were repeated continuously:

```
May  4 03:53:39 sys01 kernel: usb 1-6: new high speed USB device
using address 116
May  4 03:53:39 sys01 kernel: input: USB HID v1.01 Mouse
[Peppercon AG Multidevice] on usb-0000:00:1d.7-6
May  4 03:53:39 sys01 kernel: input: USB HID v1.01 Keyboard
[Peppercon AG Multidevice] on usb-0000:00:1d.7-6
May  4 03:53:51 sys01 kernel: usb 1-6: USB disconnect, address 116
```

Reason

The embedded SuperMicro IPMI device activated and deactivated the virtual USB mouse and keyboard for remote management in every 2 minutes causing a noticeable OS activity, which appeared also in the power consumption.

Solution

Restarting the IPMI device solved the problem by plugging off the power cord of the machine. (Reset and power off/on did not help)

As the plot shows, after reseating the power cord no noticeable system activity can be seen in idle state.

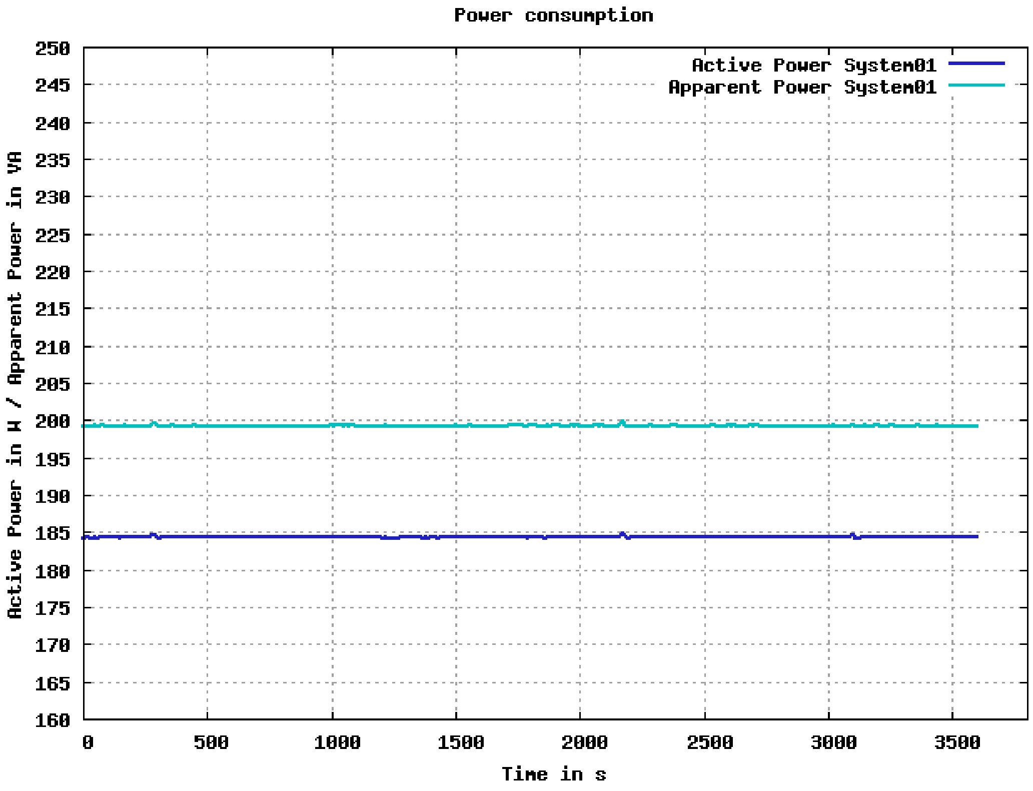


Figure 5.10 Idle results after the problem was solved

The reason of that unexpected behavior of the IPMI device is not revealed. A firmware upgrade and disabling unused virtual peripheral device functions are suggested to prevent further anomalies.

The most important reason to do this, that the regular operating system activity would most likely inhibit any deeper sleep modes on completely idle nodes.

5.4 Results

5.4.1 Tests with 2GB 667MHz FB-DIMM modules

The following charts show the results of the three main tests (idle, cpuburn, mixed cpuburn+lapack load test) conducted with identical memory configuration to compare CPUs. For the tendering process a value of 80% mixed load test Apparent power result and 20% idle Apparent power result is calculated for each configuration.

Apparent power results in VA:

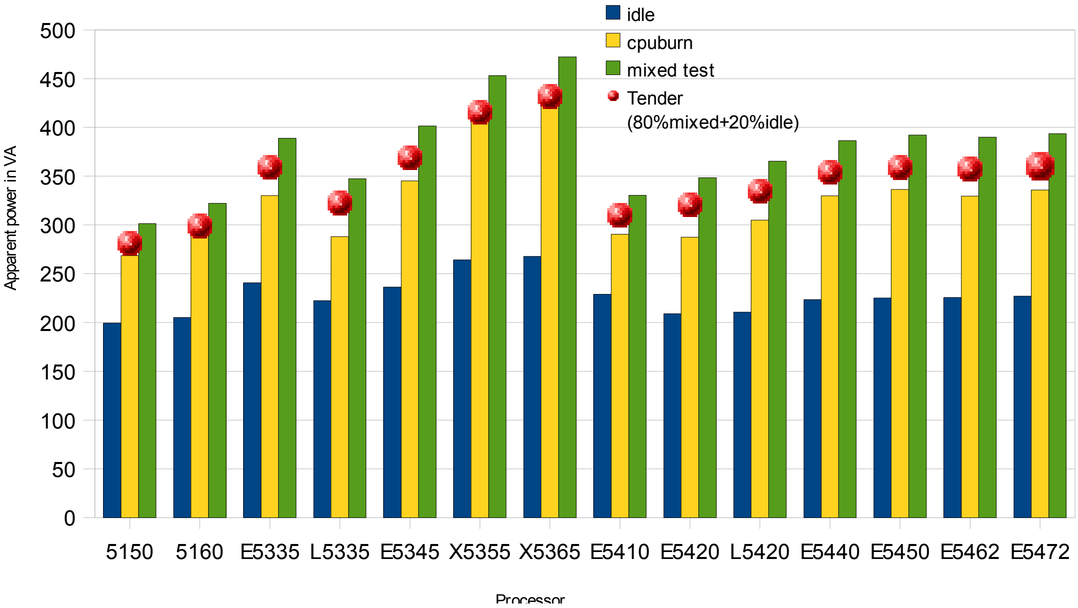


Figure 5.11 Tests with 2GB 667MHz FB-DIMM, Apparent power

Apparent power results in VA:

Family	CPU	Memory	idle	cpuburn	mixed test	Tender (80%mixed+20%idle)
Woodcrest	5150	4*2GB 667MHz	200	269	301	281
Woodcrest	5160	4*2GB 667MHz	205	292	322	299
Clovertown	E5335	8*2GB 667MHz	241	330	389	359
Clovertown	L5335	8*2GB 667MHz	222	288	347	322
Clovertown	E5345	8*2GB 667MHz	236	345	402	369
Clovertown	X5355	8*2GB 667MHz	264	422	453	415
Clovertown	X5365	8*2GB 667MHz	268	424	472	431
Harpertown	E5410	8*2GB 667MHz	229	290	330	310
Harpertown	E5420	8*2GB 667MHz	209	287	349	321
Harpertown	L5420	8*2GB 667MHz	211	305	365	334
Harpertown	E5440	8*2GB 667MHz	224	330	386	354

5.4.2 Idle test results to compare differences between memory modules

The following table shows the idle Active power (W) results of all CPU–Memory configurations.

CPU	5150	5160	L5335	E5335	E5345	X5355	X5365	E5410	E5420	L5420	E5440	E5450	E5462	E5472
Memory	8GB	8GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB
1G@667	207	212	257	273	268	297	302	245	242	248	258	261	258	257
2G@667	185	189	208	225	221	249	253	213	195	196	209	211	212	211
4G@667	171	175	181	197	193	221	225	170	167	173	182	184	184	184
Power consumption differences when using different memory module sizes but always the same amount in total.														
1-2G	22	23	49	48	48	48	49	33	47	52	48	51	46	46
2-4G	14	14	27	28	28	28	28	42	28	24	27	26	28	27
1G-4G	37	37	76	76	75	76	77	75	75	75	76	77	74	73

Even when idle, a noticeable difference can be seen in power consumption when using different memory module sizes (1GB, 2GB, 4GB). This is already an indication that it is more efficient to use a lower number of higher capacity modules.

The graph shows the results for all memory configurations for each CPU.

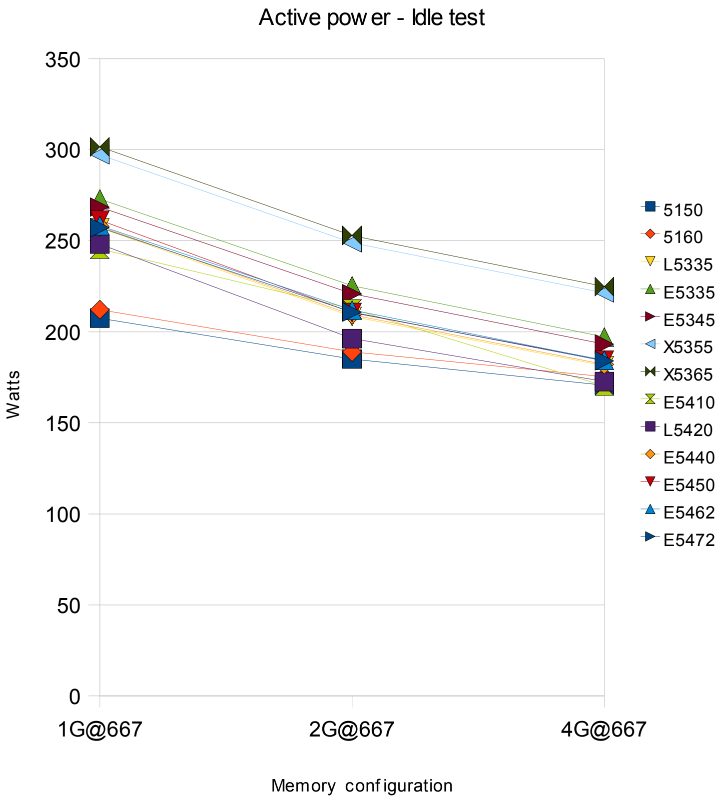


Figure 5.12 Idle test results, Active power

The changes in power consumption are consistent and almost linear with the number of modules regardless to their type.

5.4.3 Mixed CPU + Memory load test results to compare differences between memory modules

The following table shows the Active power (W) results of all CPU – Memory configurations during the mixed cpuburn+lapack tests when both the CPUs and the memory are loaded.

CPU	5150	5160	L5335	E5335	E5345	X5355	X5365	E5410	E5420	L5420	E5440	E5450	E5462	E5472
Memory	8GB	8GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB	16GB
1G@667	310	329	383	420	433	487	505	383	382	400	422	429	423	426
2G@667	289	308	335	375	388	439	459	315	336	353	374	380	378	380
4G@667	256	272	292	330	342	398	411	292	290	308	329	333	328	331
Power consumption differences when using different memory module sizes but always the same amount in total.														
1G-2G	21	21	48	45	45	47	46	67	46	46	47	49	45	47
2G-4G	33	36	42	45	46	41	48	23	46	45	45	46	50	49
1G-4G	54	57	90	90	91	88	95	90	92	91	92	95	94	95

Under load, the differences in power consumption when using different memory module sizes (1GB, 2GB, 4GB) are higher than in idle mode. Under load it is even more apparent that it is more efficient to use lower number of higher capacity modules.

The graph shows the result for all memory configuration for each CPU.

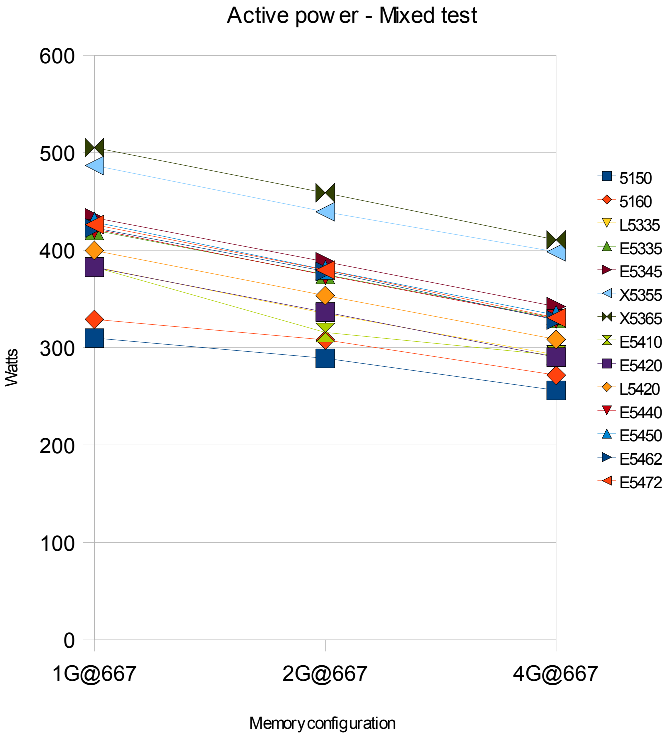


Figure 5.13 Mixed test results, Active power

The changes in power consumption are consistent, the savings in consumed power when using a smaller number of larger memory modules are quite big. Using 4GB modules instead of 1GB modules saves about 100W (about 25%) of power!

5.4.4 Comparison of 667MHz and 800MHz memory modules

The Harpertown processors with 1600MHz FSB (E5462,E5472) were tested with both 667MHz and 800MHz FB-DIMM memory modules to examine the impact of the faster frequency of the memory on power consumption.

The Active power (W) results are the following:

	Idle		Mixed	
CPU	E5462	E5472	E5462	E5472
Memory	16GB	16GB	16GB	16GB
1G@667	259	258	423	426
1G@800	258	257	429	428
2G@667	212	211	378	380
2G@800	213	213	386	389
4G@667	184	184	328	331
4G@800	191	191	339	343
Penalty for using 800MHz modules (W):				
1GB modules	-1	-1	6	2
2GB modules	2	2	8	9
4GB modules	7	7	11	12

A difference of max. ~10 Watts can be seen when using the higher frequency memory modules. The following plot shows a comparison for the system with the E5472 CPU:

Different memory setups with E5472 CPU (Active power)

	Idle 667MHz	Idle 800MHz	Mixed 667MHz	Mixed 800MHz
1G modules	258	257	426	428
2G modules	211	213	380	389
4G modules	184	191	331	343

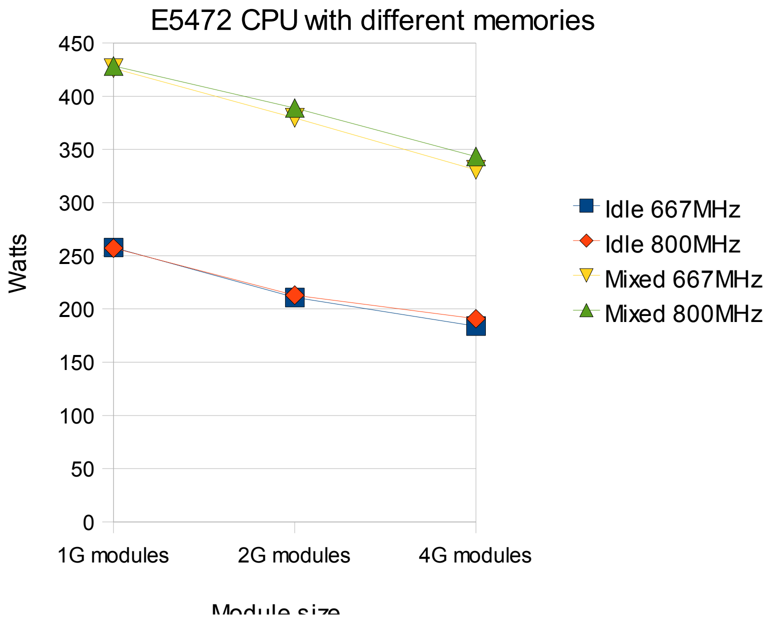


Figure 5.13 Results with different memory configurations, Active power

5.5 Analysis

Our goal with the power measurements is to get a better understanding of the CPU and memory power consumption in server computers. To achieve that, we need to calculate the amount of consumed power for the CPU and memory from the total power consumption of the machine.

For the calculation we can assume, that the machine's power consumption consists of:

- The 2 CPUs
- The Memory
- The rest of the System (motherboard, hdd, fans, etc.)

$$P_{total} = P_{cpu} + P_{memory} + P_{system}$$

5.5.1 The idle power consumption of one FB-DIMM module:

Since the idle power consumption of the system changes linearly with the number of memory modules, the energy consumed by one module can be calculated easily by subtracting the results of two different memory configurations measured with the same system:

$$P_{mem\_idle} = (P_{total\_idle\_M1} - P_{total\_idle\_M2}) / (M2\text{-}M1)$$

Where M1 and M2 are the number of modules in the actual configuration.

The calculated results:

1GB 667MHz:	6.14 W
2GB 667MHz:	5.98 W
4GB 667MHz:	5.34 W

The numbers are based on idle and cpuburn tests performed on 2 different CPU sets and 9 different memory setups for each test, and they include the additional power consumption of the chipset that is needed to drive the modules. The results show only a minor difference between the different memory module sizes.

5.5.2 The power consumption of the base system:

The power consumption of the base system contains all devices apart from the CPU and the memory after they have been separated out. The base system has a different power consumption in idle and under load when the CPU is stressed, so a number for both states have to be calculated by subtracting the CPU and the memory from the total results:

Idle system:

$$P_{sys\_idle} = P_{total\_idle} - 2*P_{CPU\_idle} - P_{memory\_idle}$$

System under load, when only the CPU is stressed:

$$P_{sys\_load} = P_{total\_cpuburn} - 2*P_{CPU\_load} - P_{memory\_idle}$$

where P_{memory_idle} is the power consumption of all memory in the computer

Only the CPU's power consumption is missing from the formulas. In order to estimate the missing numbers, several tests were done with the E5472 CPU both in idle and load using 1CPU and 2CPUs in the system. The tests were repeated also with different memory configurations. After subtracting the 1CPU results from the results with 2CPUs, the following, consistent numbers were gained:

P_{CPU_idle}:	19.4 W
P_{CPU_load}:	69.8 W

Now the base system can be calculated:

$$P_{sys\_idle} = P_{total\_idle} - 2*P_{CPU\_idle} - \#modules*P_{mem\_idle}$$

$$P_{sys\_load} = P_{total\_cpuburn} - 2*P_{CPU\_load} - \#modules*P_{mem\_idle}$$

where #modules is the number of FB-DIMM modules and P_{mem_idle} is the idle power consumption of one module.

The calculated results for the base system:

P_{sys_idle}:	123.06 W
P_{sys_load}:	132.64 W

5.5.3 Power consumption of the CPU

After examining the base system and the idle memory, the power consumption of each processor can be expressed:

$$P_{cpu}=P_{total}-P_{system}-P_{memory}$$

The calculated results can be seen in the following tables.

Idle power consumption of CPUs:

$$P_{cpu\_idle} = (P_{total\_idle} - P_{system\_idle} - \#modules * P_{mi}) / 2$$

Family	CPU	Cores	Frequency (GHz)	Idle power consumption / CPU (W)	W/GHz Idle
Woodcrest	5150	2	2,67	18,4	6,9
Woodcrest	5160	2	3	20,6	6,9
Clovertown	E5335	4	2	26,4	13,2
Clovertown	L5335	4	2	18,3	9,2
Clovertown	E5345	4	2,33	24,3	10,4
Clovertown	X5355	4	2,67	38,5	14,4
Clovertown	X5365	4	3	40,4	13,5
Harpertown	E5410	4	2,33	15,3	6,6
Harpertown	E5420	4	2,5	11,2	4,5
Harpertown	L5420	4	2,5	13,4	5,4
Harpertown	E5440	4	2,83	18,7	6,6
Harpertown	E5450	4	3	19,9	6,6
Harpertown	E5462	4	2,8	19,6	7,0
Harpertown	E5472	4	3	19,2	6,4

Load power consumption of CPUs:

$$P_{cpu\_load} = (P_{total\_cpuburn} - P_{system\_load} - \#modules * P_{mi}) / 2$$

Family	CPU	Cores	Frequency (GHz)	Load power consumption / CPU (W)	TDP (W)	W/GHz Load
Woodcrest	5150	2	2,67	49,4	65	18,5
Woodcrest	5160	2	3	59,9	80	20,0
Clovertown	E5335	4	2	67,0	80	33,5
Clovertown	L5335	4	2	47,4	50	23,7
Clovertown	E5345	4	2,33	74,7	80	32,0
Clovertown	X5355	4	2,67	113,4	120	42,5
Clovertown	X5365	4	3	114,4	120	38,1
Harpertown	E5410	4	2,33	46,8	80	20,1
Harpertown	E5420	4	2,5	46,9	80	18,7
Harpertown	L5420	4	2,5	55,6	50	22,2
Harpertown	E5440	4	2,83	67,8	80	24,0
Harpertown	E5450	4	3	71,7	120	23,9
Harpertown	E5462	4	2,8	67,6	80	24,1
Harpertown	E5472	4	3	69,6	80	23,2

In general, all CPUs stayed within their specified Thermal Design Power. Most processors consume about 20W in idle and 50-70W under full load. A very good power efficiency can be seen for all Harpertown processors especially with lower frequencies, but the faster members of the family are also very reasonable, the differences inside the Harpertown family are very low.

The highest power consumptions were measured at the X series Clovertowns, where both processor consumed exceptionally higher amount of energy for the promised higher computing capability.

The following chart shows a comparison of the Watt/GHz values for all CPUs:

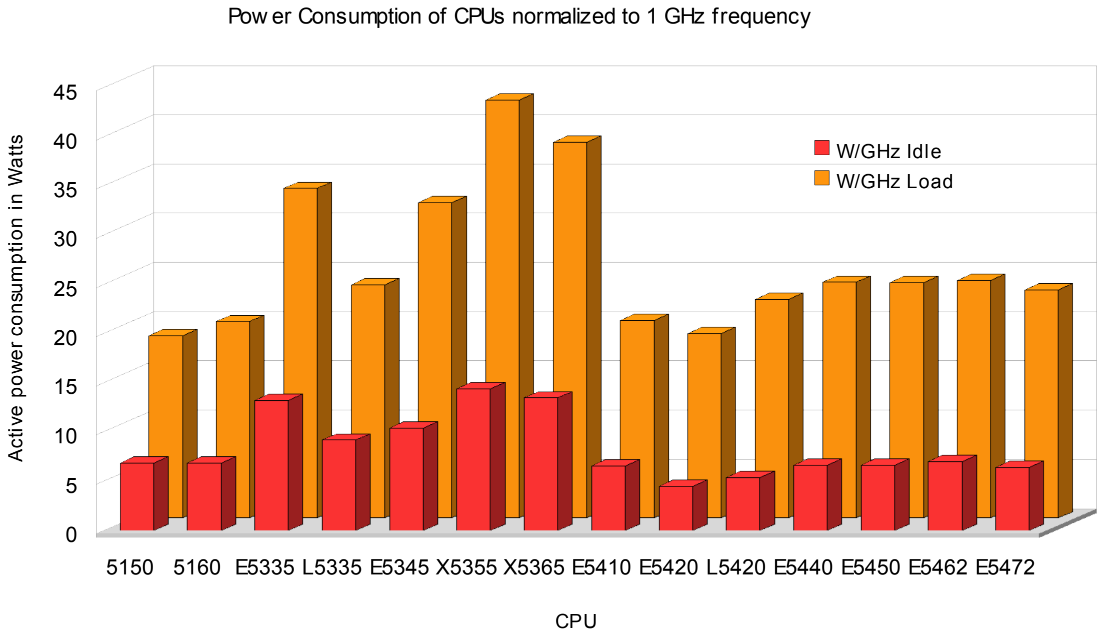


Figure 5.14 Normalized power consumption of CPUs, Active power

5.5.4 Memory load power consumption

The power consumption of the memory under load can be derived from the results of the mixed CPU+memory load tests. During this test, half of the execution cores are running lapack to generate full load on the memory and partly on the CPU. The rest of the cores are running cpuburn to stress the CPU as heavily as possible.

The load power consumption of one memory module can be calculated as follows:

$$P_{mem\_load} = (P_{total\_mixed} – P_{sys\_load} - 2*P_{cpu\_load}) / \#modules$$

The calculated results for one FB-DIMM module:

1GB 667MHz:	9.5 W
2GB 667MHz:	12.8 W
4GB 667MHz:	14.5 W

5.5.5 Power consumption of the AMB and the rest of the memory module:

The FB-DIMM memory was developed to overcome a number of limitations of DDR/DDR2 memory. The AMB was introduced to decouple the communication between the memory controller and the actual memory chips. This decoupling enabled a large number of improvement compared to DDR/DDR2 memory. Unfortunately they come at a price, because the AMB chip consumes power.

The power consumed by an FB-DIMM memory module can be separated to:

- The AMB chip
- Actual memory chips

The calculations are based on the fact, that seemingly the only difference between the tested 1GB and 2GB modules is the number of actual memory chips. Assuming that the AMB consumes the same amount of power in all modules, and assuming that the memory chips which are of the same physical and logical size consume the same amount of power on all modules, the AMB chip's power consumption can be expressed:

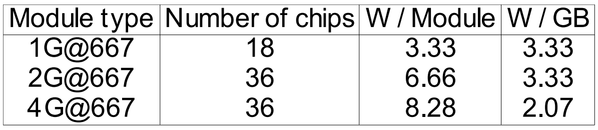
$$\text{1GB Module: } AMB + X = 9.5 \text{ Watts}$$

$$\text{2GB Module: } AMB + 2X = 12.8 \text{ Watts}$$

From these expressions the following values are gained:

- AMB Active power: 6.18 Watts

- The rest of the memory module (X), the memory chips Active power in Watts:



Module type	Number of chips	W / Module	W / GB
1G@667	18	3.33	3.33
2G@667	36	6.66	3.33
4G@667	36	8.28	2.07

5.6 Additional measurements

5.6.1 800MHz memories running at 667MHz

Several measurements were taken with configurations that are able to use the 800MHz memory modules only at a speed of 667MHz to compare the results between the normal 667MHz modules and the 800MHz modules that are running at 667MHz. The following tables show the results:

Idle				
CPU	5160	L5335	E5335	E5345
Memory	16GB	16GB	16GB	16GB
1G@667	262	257	273	268
1G@800 (667)	252	247	263	259
4G@667	186	181	197	193
4G@800 (667)	190	185	201	
Penalty for using 800MHz modules running on 667MHz (W):				
1GB modules	-10	-10	-10	-10
4GB modules	4	4	4	

Mixed test				
CPU	5160	L5335	E5335	E5345
Memory	16GB	16GB	16GB	16GB
1G@667	397	372	409	422
1G@800 (667)	386	383	420	433
4G@667	307	292	330	342
4G@800 (667)	309	294	331	
Penalty for using 800MHz modules running on 667MHz (W):				
1GB modules	-11	11	11	11
4GB modules	2	1	1	

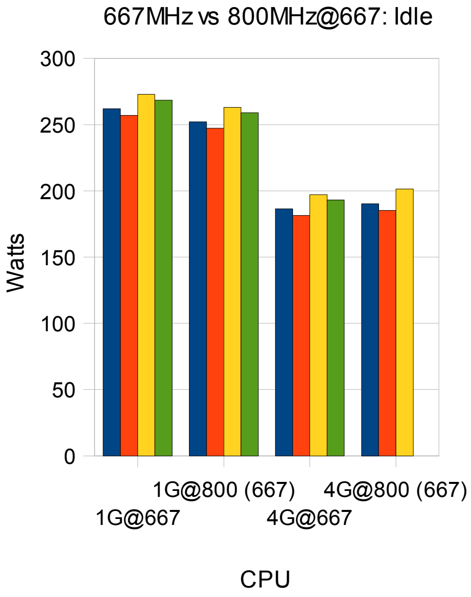


Figure 5.15 FB-DIMM comparison when Idle, Active power

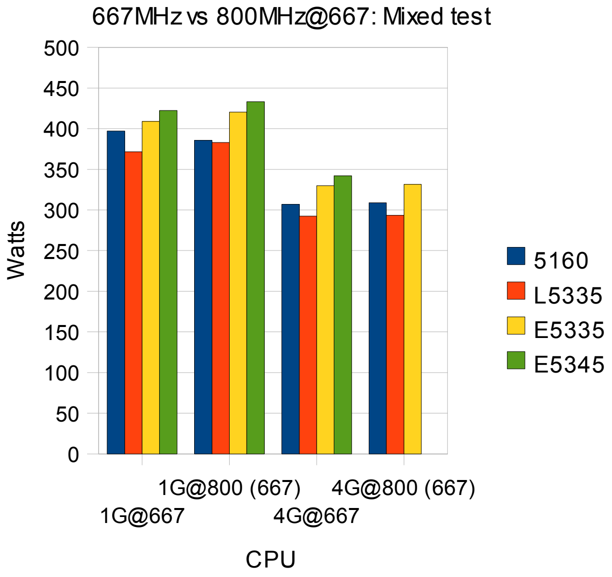


Figure 5.16 FB-DIMM comparison under Load, Active power

It can be seen on the charts, that in some cases a power saving of ~10W is possible in idle mode with the 1GB modules. But this gain is lost under load. The results for the 5160 Woodcrest CPU show an exceptional behavior, 10-11W was saved in both idle and load by using 800MHz modules. That particular test was repeated to assure the validity of the results. When using 4GB modules, in all cases the configuration with 800MHz memory consumes 1-4W more than the one with 667MHz.

5.6.2 E5420 and L5420 comparison

The comparison of the E5420 and L5420 CPUs brought unexpected results, so additional tests were performed with these CPUs.

The L5420 Harpertown CPU is of the power efficient LV series of Intel Xeon CPUs which means a lower TDP than the standard E5420. The TDP of the L5420 is 50W, which is significantly less than the 80 Watts of the E5420. However the results showed no noticeable difference between the two processors, or even the L type CPU consumed more. The explanation for this behavior is, that the actual power consumption of a given CPU can, of course, be significantly lower than the stated TDP. In some cases a standard E type processor can therefore have a lower power consumption than an LV classified processor.

Apparent power results for systems with the two CPUs:

	E5420	**L5420**
Idle	209 VA	211 VA
Load (mixed)	349 VA	365 VA

To verify the values another set of test was performed on an evaluation system from a different vendor (Dell) to avoid all possible anomalies caused by the current base system.

The Apparent power results on the Dell PowerEdge 1950 evaluation system:

	E5420	**L5420**
Idle	206 VA	206 VA
Load (mixed)	346 VA	348 VA

There is no significant difference between the results on the two systems for the E5420. The higher power consumption of the L5420 in the standard test system is unexpected. There is no explanation for this behavior. One possible explanation could be the different chipsets.

Another question arises from the fact that the power consumption of the same model CPUs can show such large differences. Is the financial penalty after power consumption provides an accurate assessment for the tendering process?

Let's say, that a distributor offers 1000 servers with an average of 500W power consumption. Then the one server with the lowest power consumption is being sent for evaluation by chance. The performance tests are made, the power consumption is measured, and the distributor wins the tender with the low-consumption evaluation system. Now after paying the bill, the 1000 server arrives, and after the final installation, the power consumption is measured again, but this time also on the “average” systems. The power consumption is of course higher than what was measured at the evaluation, but who is the responsible?

Unfortunately there is no solution yet for this problem.

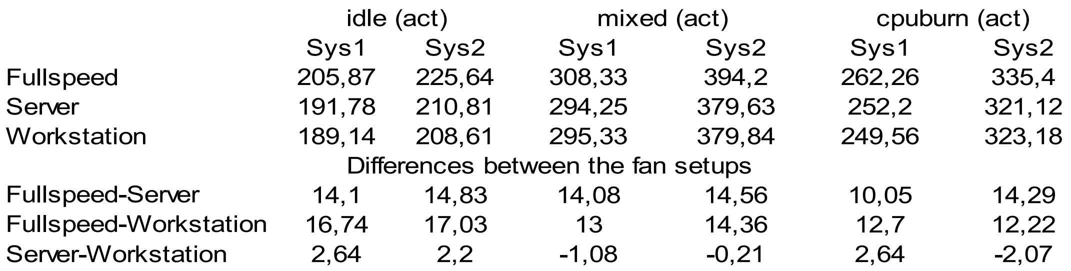
5.6.3 Fan settings

A series of tests were performed on two hardware configurations to see the impact on power consumption of the different BIOS controlled chassis fan setups.

The BIOS of the SuperMicro X7DWN+ motherboard provides 3 options to set up the fan's behavior:

- Workstation
- Server
- Full speed

Idle, cpuburn and mixed lapack+cpuburn tests were run in all modes, the results are shown in the following table:

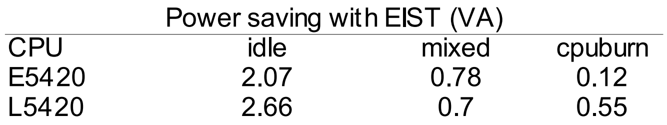


	idle (act)		mixed (act)		cpuburn (act)	
	Sys1	Sys2	Sys1	Sys2	Sys1	Sys2
Fullspeed	205,87	225,64	308,33	394,2	262,26	335,4
Server	191,78	210,81	294,25	379,63	252,2	321,12
Workstation	189,14	208,61	295,33	379,84	249,56	323,18
	Differences between the fan setups					
Fullspeed-Server	14,1	14,83	14,08	14,56	10,05	14,29
Fullspeed-Workstation	16,74	17,03	13	14,36	12,7	12,22
Server-Workstation	2,64	2,2	-1,08	-0,21	2,64	-2,07

There is no noticeable difference between server and workstation mode, but the computers in full speed mode consume 14 watts more during all tests. This means that in either server or workstation mode the fans never run at full speed during any of the tests. A result like this is expected, since the temperature in the lab-room where the measurements were done was low and relatively constant. In a warmer environment the fans would certainly have to run at higher speed and therefore consume more power, and the difference between the Server and Workstation mode would also be more significant.

5.6.4 Tests with enabled EIST

A test with two sets of CPUs was performed to see the possible power savings that can be reached by enabling the Enhanced Intel SpeedStep Technology option in the computer's BIOS. The EIST capability provides dynamic frequency and core voltage scaling in order to save power when the system is idle.



Power saving with EIST (VA)

CPU	idle	mixed	cpuburn
E5420	2.07	0.78	0.12
L5420	2.66	0.7	0.55

The results unexpectedly showed only about 1% less power consumption in idle and there were no significant difference under load.

5.6.5 Test with compiler optimization

A test was performed to examine what is the impact of the different compiler options on the power consumption. Lapack tests were run with different gcc compiler options in the following order:

- Standard
- -m32
- -m32 -O2
- -O2

The -O2 parameter means 2nd level compiler optimization, -m32 means 32bit compatible mode.

The following graph is the plot of the gained results:

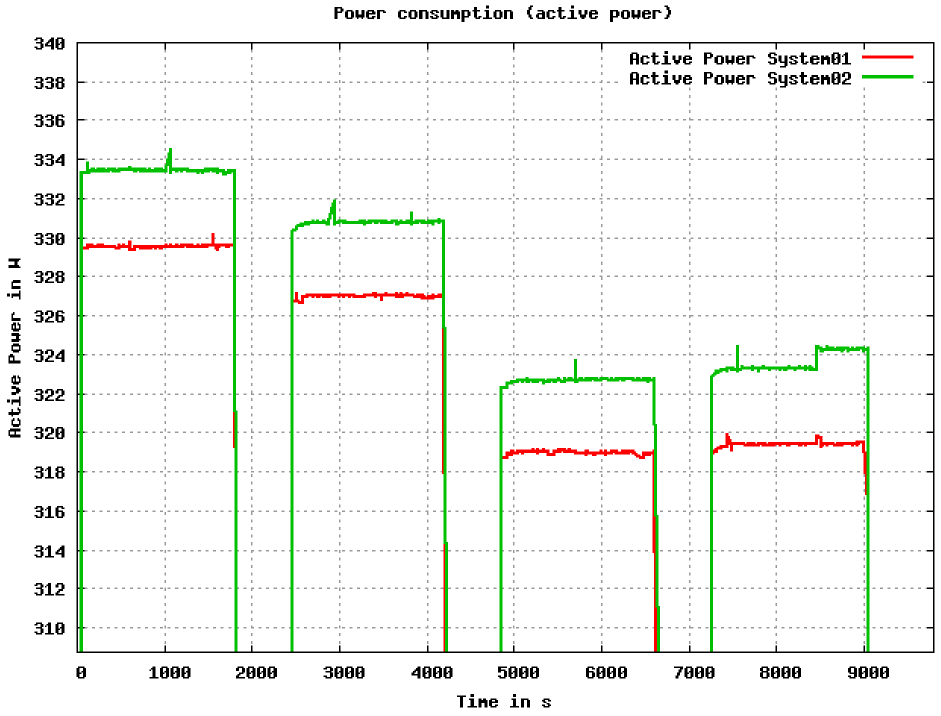


Figure 5.17 Results with compiler optimization, Active power

The plot shows a noticeable ~4% difference when the compiler is optimized.

When the 32 bit mode is used, the program doesn't use all the available resources, and that can be seen in both performance and power consumption.

5.6.6 Blade solutions

As the alternative of traditional computers, the high density blade solutions were also tested to see how efficiently they work compared to standard density servers. Power measurements were performed on two enclosures. One enclosure contained 16 blades from the HP b1416 series, and the other was an HP Bl2x220c double density solution containing 32 nodes. In both enclosures, the blades were dual-socket systems equipped with quad-core Harpertown processors.

The apparent power figures for the double density blades showed quite large improvements in power consumption compared to the 4U system with identical CPUs. There were 50% less power consumption when idle, and 33% improvement under load. The improvement is at least partly related to the type of the memory. The double density blades are equipped with standard unbuffered DDR2 memory, while the comparison 4U test systems, as well as the B1416c blades are equipped with FB-DIMM modules. The B1460c systems showed no difference in idle, and only 4-5% improvement under load compared to the 4U test systems. Taking into account, that the 4U systems have 2 SATA disks and a bit less efficient power supply, the difference is even smaller, if not negligible.

The power factor was always more than 0,97 in idle, and more than 0.99 under load in all blades, which also proves that larger power supplies tend to be more effective.

5.6.7 The value of low-cost, low-power processors: Intel Atom N330 [5]

As one of the possible future directions in computing, the value of low-cost, low-power processors were examined recently. A 'home-built' single socket server equipped with Intel's recently released dual-core Atom N330 processor was compared against a dual-socket Xeon Core 2 Quad server in both performance and power consumption. The throughput of the 1.6GHz Atom system is of course far below that of the 3GHz Harpertown system, but the price and power consumption difference makes the Atom a considerable alternative. The ratio in throughput was 13.3 in favor of the Harpertown server, but comparing the prices on the web, the Atom is almost 20 times cheaper. The Atom consumes 5 times less power than the dual-socket server, which is less impressive than expected considering the 8W power consumption of the Atom N330 CPU. Probably the chipset is responsible for the relatively high power consumption of the Atom system (50.7 W), but the figures are already very promising. Currently the Atom is no match for the Xeon based servers in a power constrained environment, but the idea of using low-cost, low-power processors in computing seems to be a considerable alternative in the future.

5.7 What to do when buying a thousand servers? Conclusion of the measurements

The results of the power measurements revealed a very high possibility of large power savings in server computers by choosing the appropriate CPU and memory configuration. The difference in power consumption between systems that provide similar performance can reach 150W (almost 50%) comparing a power saver configuration and a setup with higher power consumption.

Choosing the right CPU is a very important fact to save power. More than 40 Watts can be saved on each GHz in a dual-socket system comparing the best and the worst case. In general, processors with more cores and lower feature size tend to be more power efficient. In the current test, both the lower and higher frequency members of the Harpertown family CPUs provided excellent power efficiency.

However it is very problematic to set up a reliable evaluation method for processors, since there is no way at the moment to tell the actual power consumption of a CPU without measurement. Only the TDP is given by the manufacturer which guarantees the maximum limit of power that is used by the CPU, but the actual power consumption can vary between CPUs of the same type.

The measurements have confirmed that the power consumption of today's main memory can be compared with that of the processor, and in some cases the memory can consume even more energy than the CPU. Therefore it is also very important to consider the power efficiency when choosing the type of the main memory.

A very high amount of energy can be saved by using a lower amount of higher capacity memory modules. Building a system from 4GB FB-DIMM modules instead of using 1GB modules can save 90 Watts on the power consumption in an 8 core system equipped with 16GB memory. The modules with different frequencies have also been compared, and the result indicated only a minor difference in power consumption between the tested 667MHz and 800MHz FB-DIMM modules.

The detailed analysis of the FB-DIMM modules pointed out that the Advanced Memory Buffer (AMB) chip consumes a very high amount of power (~6.2W) on each module. Although FB-DIMMs provide enhanced capabilities over unbuffered DDR/DDR2 modules, the large difference in power consumption caused by the AMB chip has to be considered when building power efficient server computers. At the moment, configurations using Intel's San Clemente (5100) chipset with unbuffered DDR2 memory can offer remarkable advantage in power efficiency. DDR3 will be more available in the future, providing additional power savings and enhanced performance.

Blade solutions are introducing increased density, and the shared facilities, such as the power supplies, cooling and monitoring are expected to increase also the power efficiency. The power measurements indicated reasonable power consumption for the tested double density blade solution, but the results also showed, that the power consumption of blades can also be unimpressive compared to standard form factor computers, while the highly increased density may come with powering and cooling difficulties.

As another possible direction of future developments, the value of low-cost, low-power CPUs were tested as well, and with their reasonable throughput/power/price ratio, they were found comparable even with the recent Quad-Core processors. The future generations of low-cost, low-power CPUs may become a considerable alternative for specific purposes in power constrained environments.

The results of power measurements indicated that there are several other possibilities for additional, generally minor power savings by using the appropriate settings in BIOS e.g. for the fan speed control. The review of settings on the current computers in production may lead to a better utilization of the available resources without any further investment.

6. Valuation, possible improvements

The document discussed the power consumption problems of data centers and possible solutions to improve overall power efficiency. The main subject of the investigation was how power measurements can contribute in decreasing power consumption.

The several power measurements that were performed on different configurations showed very large differences even between computers that are very close in performance. The necessity for measuring power is unquestionable, especially because power consumption under different, user specific loads may also vary.

The performed measurements revealed a very high possibility of power savings by choosing the appropriate components when building an HPC (High Performance Computing) server system. In general, processors with lower feature size and more cores are significantly more efficient than their predecessors, and also a lot of energy can be saved by using less memory modules of larger capacity.

The large amount of results made it possible to calculate the actual power consumption of the main components inside the computer, although the power was measured outside on the power cord. These estimates will give a basis for further comparisons and support the procurement team in identifying the most considerable power saver configurations during the acquisitions.

A possible improvement would be, to take computing performance into account when analyzing the results of the power measurements. In the current document it was assumed, that higher frequency processors provide more throughput.

At the moment, the final throughput per watt figures are calculated during the acquisition based on two different measurements. Once the power is measured, and an average is calculated by taking 80% load and 20% idle power consumption, and then the throughput is measured separately with the SPEC2006 benchmark suite.

An ideal case would be, if a benchmark could be developed, which would simulate the actual HEP (High Energy Physics) applications, and would measure HEP specific throughput, while the power measurements performed during the test would estimate the power consumption of the computer in full production. There are no such benchmark found yet on the market, the current benchmarks are often measure different type of throughput, and because of high diversity in power consumption during the tests, they are not suited for power measurements. Currently a set of C++ benchmarks from the SPEC2006 benchmark suite is used to measure performance.

As an alternative solution, it would be possible to examine the relation between the power consumption figures under the used mixed CPU+memory load test, and under the actual HEP applications. If there is a constant ratio between the values, more accurate power consumption estimations will be allowed.

7. References

[1] Dr. Andreas Hirstius - Sverre Jarp: Strategies for increasing data centre power efficiency, CERN openlab Mar 2008
http://openlab-mu-internal.web.cern.ch/openlab-mu-internal/Documents/2_Technical_Documents/Technical_Reports/2008/AH-SJ_The%20approach%20to%20energy%20efficient%20computing%20at%20CERN%20final.pdf

[2] The Green Grid: Data Center Power Efficiency Metrics: PUE and DciE 2007
http://www.thegreengrid.org/gg_content/TGG_Data_Center_Power_Efficiency_Metrics_PUE_and_DCiE.pdf

[3] Margery Conner: Push for power efficiency forces changes in server-center hardware and software EDN 01 Jun 2008
http://www.edn.com/article/CA6541394.html

[4] Next Generation Power & Energy: Accurately Measuring Data Center Power Efficiency, 03 Oct 2008
http://www.nextgenpe.com/pastissue/article.asp?art=272218&issue=231

[5] Gyorgy Balazs, Sverre Jarp, Andzej Nowak: Is the Atom ready for High Energy Physics? - CERN, Nov 2008
http://openlab-mu-internal.web.cern.ch/openlab-mu-internal/Documents/2_Technical_Documents/Technical_Reports/2008/CERN%20Atom%20330%20analysis.pdf

[6] SprayCool Inc.: IGBT white paper, 2007
http://www.spraycool.com/products/IGBT%20cooling%20r1.pdf

[7] Wikipedia: CERN, 08 Dec 2008
http://en.wikipedia.org/wiki/Cern

[8] 42U: Improve energy efficiency, reduce power consumption, lower energy costs
http://www.42u.com/power/data-center-power.htm , 12 Dec 2008

[9] Wikipedia: AC power, 18 Oct 2008
http://en.wikipedia.org/wiki/AC_power

[10] Mag Securs: Datacenter Energy Costs on the Rise, 02 Oct 2008
https://www.mag-securs.com/spip.php?article11743

[11] Bob Worrall: A Green Budget Line, 28 Jul 2008
http://www.forbes.com/technology/2008/07/27/sun-energy-crisis-tech-cio-cx_rw_0728sun.html

Appendix: The power measurement process

The following short note describes how the measurements are being done from starting the power meter to filling up the data sheets.

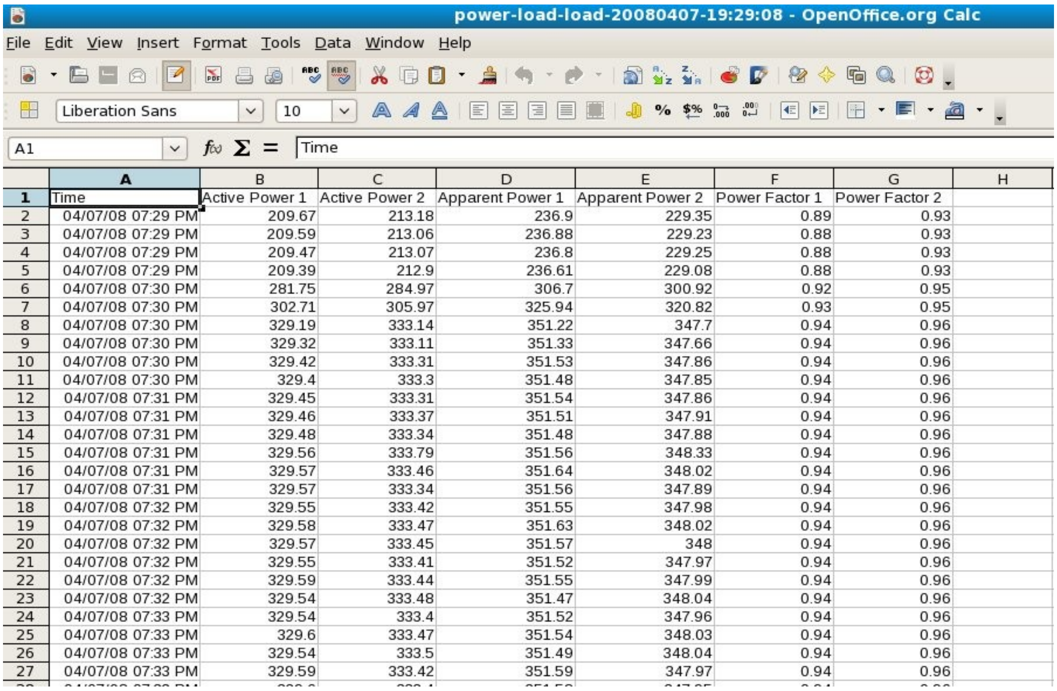
Starting power measurement manually on the power meter laptop

A simple power measurement can be started on the laptop connected to the power meter by running /home/power/runpower.sh with the necessary parameters. We need to declare:

- The type of test (load/idle)
- The name of the tested machine
- Directory to store the results
- Power meter channels

After starting the measurement, the power meter writes out the actual power consumption values for all specified channels in a single CSV file, which will be stored in /home/power/results/pcpowerm on the power meter laptop.

Command: /home/power/runpower.sh -t load -m testmachine -r /home/power/results/pcpowerm -x"-c 1,2" -l 410m >> measurements.log


power-load-load-20080407-19:29:08 - OpenOffice.org Calc
File Edit View Insert Format Tools Data Window Help
Liberation Sans
10
A1
Time

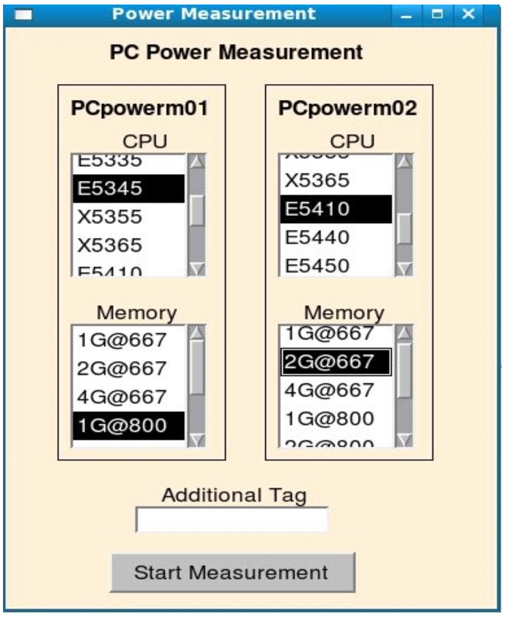
	A	B	C	D	E	F	G	H
1	Time	Active Power 1	Active Power 2	Apparent Power 1	Apparent Power 2	Power Factor 1	Power Factor 2	
2	04/07/08 07:29 PM	209.67	213.18	236.9	229.35	0.89	0.93	
3	04/07/08 07:29 PM	209.59	213.06	236.88	229.23	0.88	0.93	
4	04/07/08 07:29 PM	209.47	213.07	236.8	229.25	0.88	0.93	
5	04/07/08 07:29 PM	209.39	212.9	236.61	229.08	0.88	0.93	
6	04/07/08 07:30 PM	281.75	284.97	306.7	300.92	0.92	0.95	
7	04/07/08 07:30 PM	302.71	305.97	325.94	320.82	0.93	0.95	
8	04/07/08 07:30 PM	329.19	333.14	351.22	347.7	0.94	0.96	
9	04/07/08 07:30 PM	329.32	333.11	351.33	347.66	0.94	0.96	
10	04/07/08 07:30 PM	329.42	333.31	351.53	347.86	0.94	0.96	
11	04/07/08 07:30 PM	329.4	333.3	351.48	347.85	0.94	0.96	
12	04/07/08 07:31 PM	329.45	333.31	351.54	347.86	0.94	0.96	
13	04/07/08 07:31 PM	329.46	333.37	351.51	347.91	0.94	0.96	
14	04/07/08 07:31 PM	329.48	333.34	351.48	347.88	0.94	0.96	
15	04/07/08 07:31 PM	329.56	333.79	351.56	348.33	0.94	0.96	
16	04/07/08 07:31 PM	329.57	333.46	351.64	348.02	0.94	0.96	
17	04/07/08 07:31 PM	329.57	333.34	351.56	347.89	0.94	0.96	
18	04/07/08 07:32 PM	329.55	333.42	351.55	347.98	0.94	0.96	
19	04/07/08 07:32 PM	329.58	333.47	351.63	348.02	0.94	0.96	
20	04/07/08 07:32 PM	329.57	333.45	351.57	348	0.94	0.96	
21	04/07/08 07:32 PM	329.55	333.41	351.52	347.97	0.94	0.96	
22	04/07/08 07:32 PM	329.59	333.44	351.55	347.99	0.94	0.96	
23	04/07/08 07:32 PM	329.54	333.48	351.47	348.04	0.94	0.96	
24	04/07/08 07:33 PM	329.54	333.4	351.52	347.96	0.94	0.96	
25	04/07/08 07:33 PM	329.6	333.47	351.54	348.03	0.94	0.96	
26	04/07/08 07:33 PM	329.54	333.5	351.49	348.04	0.94	0.96	
27	04/07/08 07:33 PM	329.59	333.42	351.59	347.97	0.94	0.96	

A.1 Raw data CSV file from measurement using 2 channels

Performing the test from a remote workstation

The test process can be controlled remotely from any workstation by the standard_test.sh script, which uploads and starts the tests on the appropriate machines, and also controls the power meter at the same time.

The particular tests are situated in the power/run-measurements/#testname# directory, and they are uploaded and executed by the standard_test.sh script.



A.2 Startgui.py -graphical interface for the standard_script.sh

Directory hierarchy

The generated results are stored in the following directory hierarchy:

Power

- Results
    - All
    - #CPU
        - #Memory
            - Idle test CSV file
            - Load test CSV file
            - Memory configuration
            - CPU configuration
- Plots
    - All
    - #CPU
        - #Memory
            - Idle test active/apparent power plot PNG file
            - Idle test power factor plot PNG file
            - Load test active/apparent power plot PNG file
            - Load test power factor plot PNG file
- Sheets
    - CPUsheets
        - #CPU spreadsheet file
    - Testsheets
        - #Test spreadsheet file