University of Castilla–La Mancha

A publication of

Department of Computing Systems

Power consumption of HPC applications

by

J.L. Sánchez, F.J. Alfaro, R. Galindo,
M. Alonso, S. Coll, P. López, J.M. Rubio
F.J. Andújar

Technical Report #**DIAB-21-02-1** January 2021

(*) This work has been jointly supported by the MINECO and European Commission (FEDER funds) under the project TIN2015-66972-C5-2-R, Junta de Comunidades de Castilla-La Mancha under the project SBPLY/17/180501/000498, and Intel Coorporation under the agreement CONV180068.

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS
ESCUELA SUPERIOR DE INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE CASTILLA–LA MANCHA
CAMPUS UNIVERSITARIO s/n
02071, ALBACETE, SPAIN
Phone +34.967.599200, Fax +34.967.599224

Power consumption of HPC applications

José L. Sánchez, Francisco J. Alfaro, Raúl Galindo
Computing Systems Department
Faculty of Computer Science Engineering
University of Castilla-La Mancha
02071 – Albacete, Spain
{jose.sgarcia,fco.alfaro,raul.galindo}@uclm.es

Marina Alonso, Pedro López, Salvador Coll, Juan M. Martínez
Department of Computer Engineering
Universitat Politècnica de València
46022 – Valencia, Spain
{malonso,plopez}@disca.upv.es, scoll@eln.upv.es,jmmr@upv.es

Francisco J. Andújar-Muñoz
Computing Systems Department
Faculty of Computer Science Engineering
University of Valladolid
47011 – Valladolid, Spain
fandujarm@infor.uva.es

Contents

Abstract

HPC systems are significant energy consumers, and achieving an optimal performance/energy ratio is a challenge for massively parallel computer architects. Efforts to design energy-efficient compute nodes have been underway for a long time. However, much lower attention has been paid to the interconnection network energy consumption.

Proposals to reduce consumption of interconnection systems are tested and evaluated through simulation and the use of consumption models. These models require power consumption data for the different components of the interconnection network. There are numerous documents that provide data on the power consumption of the main components of an HPC system or of the system as a whole. However, these data, for the same type of components, offer significant differences, and thus it is not easy to choose one in particular. Therefore, we have decided to collect power consumption data in a real system, using a meter for this purpose. In this report, we include all the results obtained by measuring the consumption of a cluster when HPC applications are running on it.

1 Introduction

High Performance Computing (HPC) systems are composed of hundreds or thousands of compute nodes, and provide the best possible computing support for computational problems not addressable by accessible commercial computers. All these computing elements communicate to exchange information through an interconnection system, also formed by a large number of components.

Because of this large number of elements that work uninterruptedly to run applications that require days or weeks to deliver results, HPC systems are significant energy consumers. Current most powerful supercomputer, Fugaku, reaches 28.3 MW. This is high enough to supply energy to more than 33,000 homes, according to U.S. Energy Information Administration standard. The most-energy efficient system, MN-3, achieves 21.1 GFlops/W. A projection of that metric to an exascale supercomputer with equivalent efficiency predicts power consumption will peak 473 MW, soon reaching gigawatt figures. In order to mitigate this trend, energy-proportional compute nodes are being deployed. The strategy is modulating energy consumption at server level according to processors utilization.

As the consumption of the computation nodes is reduced, the contribution of the interconnection system to the total consumption of the HPC system becomes increasingly important, ranging from 12% to 50% depending on servers utilization (higher budget for lower server utilization) [2, 6].

Therefore, new proposals to reduce consumption of interconnection systems are required. These proposals are generally tested and evaluated through simulation and the use of consumption models. These models require power consumption data for the different components of the HPC system. There are numerous documents that provide these data. However, these data, for the same type of components, offer significant differences, and thus it is not easy to choose one in particular.

We have decided to collect power consumption data in a real system, using a meter for this purpose. In this report, we include all the results obtained by measuring the consumption of a cluster when HPC applications are running on it. This data can help determine the most appropriate values for our consumption model in particular, and any other in general.

The rest of the document is organized as follows. In Section 2 we describe the main characteristics of the device used to carry out the power consumption measurements. Section 3 presents the testbed, including HPC system characteristics and parallel applications. Then, in Section 4, we present all the results obtained, and finally, in Section 5 we outline the conclusions and future work.

2 Power analyzer

We have used the Yokogawa PZ4000 power analyzer [4, 7] to measure the power consumption of the nodes and the switch in the cluster. This instrument has three operation modes: measurement of transient power, harmonic analysis and measurement of power in a period, which is the mode used in our case. Some of the most relevant characteristics this instrument can provide are included in Table 1.

Table 1: Basic characteristics of the PZ4000 power analyzer.

Accuracy	0,1% of the obtained read + 0,025% of the DC rank up to 2 Mhz
Wide bandwidth	DC up to 2Mhz
Smapling	Up to 5 MS/s
Input voltage	30 Vpeak up to 2000 Vpeak
Inout current	100 mApeak up to 20 Apeak

Figure 1 shows the block diagram of PZ4000 power analyzer. This kind of analyzers can incorporate up to four electrically isolated modules, the voltage and current signals of each one being normalized through an operational amplifier and sent to an A/D converter and to a zero passage detector circuit. The available model has four modules. The voltage and current signals are sampled at a frequency that allows the acquisition of up to 5MS/s (megasamples per second). The outputs of both blocks are transferred to the memory section through photo isolators, protecting the meter from possible electrical failures of the input sensors. The standard memory size is 100Kwords, expandable to 4Mwords. The CPU/DSP section reads the stored data and calculates the measurements such as peak values, cosine(α), active power, etc. From this block the information is stored on disk or transmitted using a GPIB or RS232 interface to a personal computer for further processing.

In our cluster, the analyzer is placed between the power supply and nodes and switch. We have used the RS232 connection to configure the analyzer and to collect data of power consumed every two seconds in three nodes as well as in the switch.

3 Test bed configuration

The test bed is composed of one switch Mellanox SB7800 Series 2 EDR 100Gb/s with 36 ports and six compute nodes HPE ProLiant DL380 Gen10 Server. Each node has two Intel Xeon Silver 4116 processors, each one with 12 cores. The operating system is CentOS 8 running OpenSM 3.3.19.

We have conducted several tests using the following parallel applications:

- *Graph500 benchmark* using the *replicated-csr* implementation, a scale factor of 20 and an edge factor of 16 [1]. All the communications are generated by MPI col-

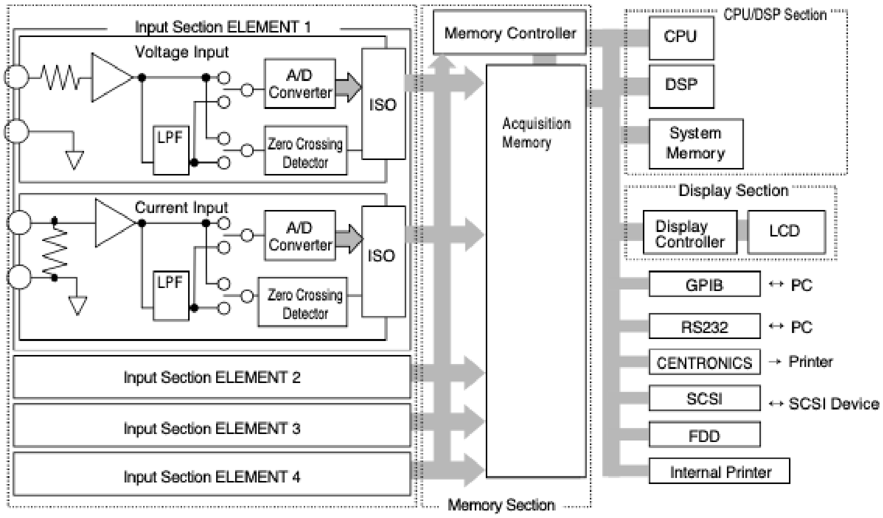


Figure 1: Block diagram of PZ4000.

lective communications that generate a great exchange of data among tasks. This application generates the highest network load of the three applications tested.

- *HPCC Linpack* [3] is used to solve a dense system of linear equations. This application follows a specific pattern in which a specific task use to communicate always with the same tasks, following a ping-pong traffic pattern.
- *HPCG* HPCG is intended as a complement to the Linpack benchmark, in order to match a different and broad set of important applications (www.hpcg-benchmark.org).
- *Gromacs* [5] is a scientific application to perform molecular dynamics. It shows a great spatial locality. We have used some of benchmarks available in the Gromacs benchmark[1].

The tests have consisted in launching processes on the nodes, varying the number of processes launched to each node, the size of the problem, as well as the number of nodes involved in each test. For each test, 30 runs have been made under the same conditions and the average power consumption measures have been taken, also considering the confidence intervals at 95% confidence level. The sampling time is 2 seconds and between each run of the application a period of time of 20 seconds (10 samples) is left to synchronize the begining and end of each execution.

[1] `http://www.gromacs.org/About_Gromacs/Benchmarks`

4 Power consumption results

This section includes the power consumption results obtained from the various tests carried out. These results are organized in several sections, namely compute nodes, switch and network card.

4.1 Node power consumption

Figures 2, 3, 4 and 5 show power consumption data for applications Graph500, Linpack, HPCG[2], and Gromacs, respectively. Sub-figures (a) show results obtained when the applications have been executed using only one node and the number of tasks has been varied from 1 to 48. As you can see power consumption increases linearly with the number of tasks up to 24 tasks (one per core), and the increase is much smaller from that number onwards. Note that there are 24 physical cores per node, but using hyperthreading 48 tasks have been launched. It can also be seen that the power consumption of an idle node is approximately 100 watts, and that when the application runs with only one task, only one core works, the power consumption is 150 watts.

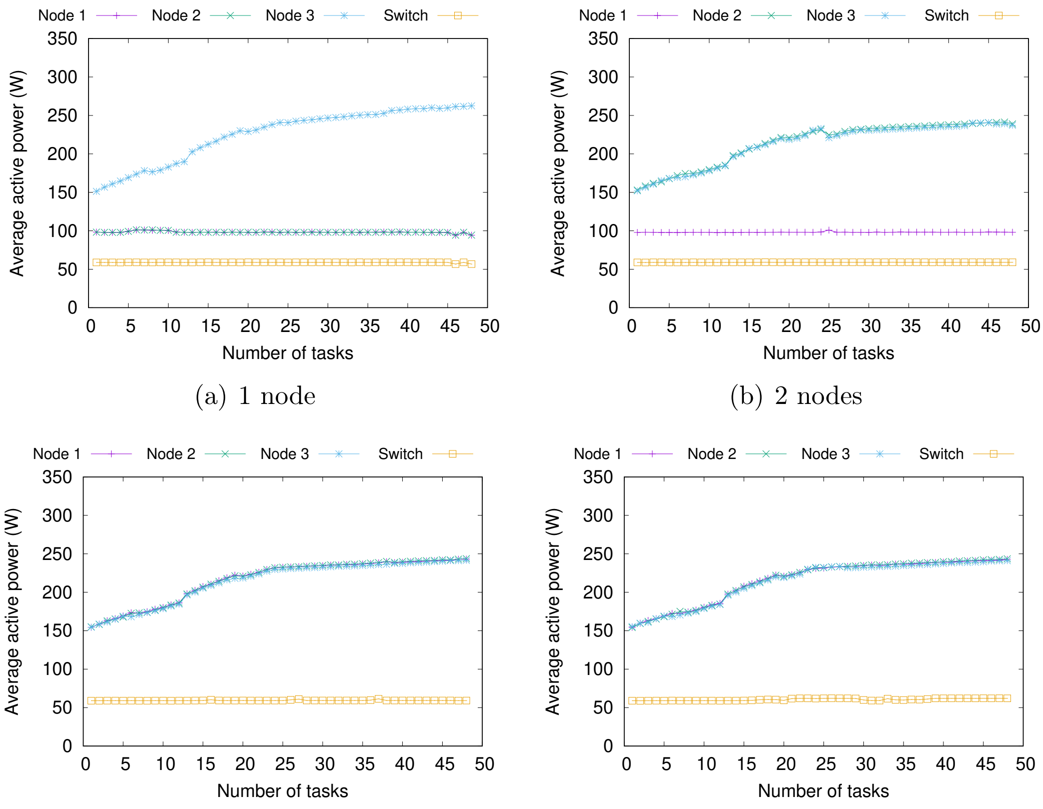
Sub-figures (b), (c), and (d) show power consumption results when the applications run on 2, 3 and 6 nodes, respectively.

On the other hand, Figures 6, 7, 8 and 9 show the results obtained when the applications have been executed considering 48 tasks per node. It can be observed that the power consumption achieved by the compute nodes is higher when the application runs only on one node. When several nodes are used, the consumption is slightly lower, and this is true regardless of the number of nodes used, in almost all cases.

It is also clear from the figures that the three nodes in the graphs behave in the same way. We can even say that the six nodes have the same behaviour because several tests have been repeated connecting the other three nodes to the power meter and the results have been the same.

Figures 10, 11, 12 and 13 show the power consumption over time. Once the initialisation stage of each application is exceeded, it can be seen how the power consumption remains practically constant.

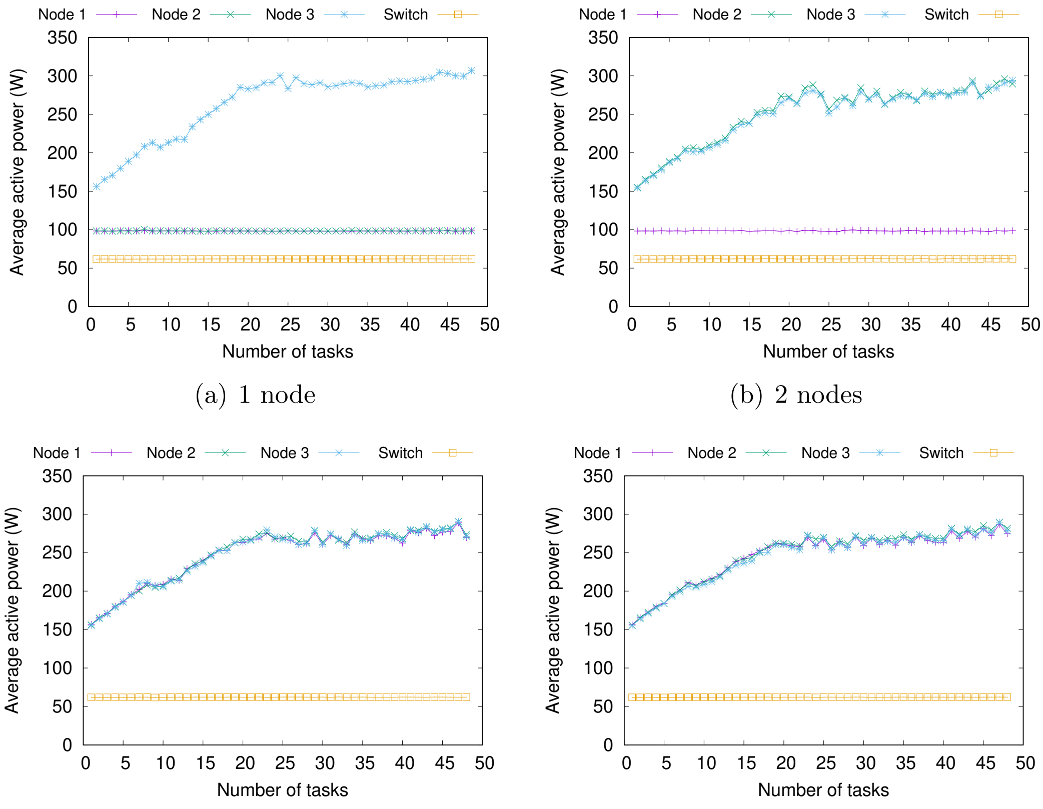
[2] HPCG is intended as a complement to the Linpack benchmark, in order to match a different and broad set of important applications (www.hpcg-benchmark.org).



(a) 1 node (b) 2 nodes

(c) 3 nodes (d) 6 nodes

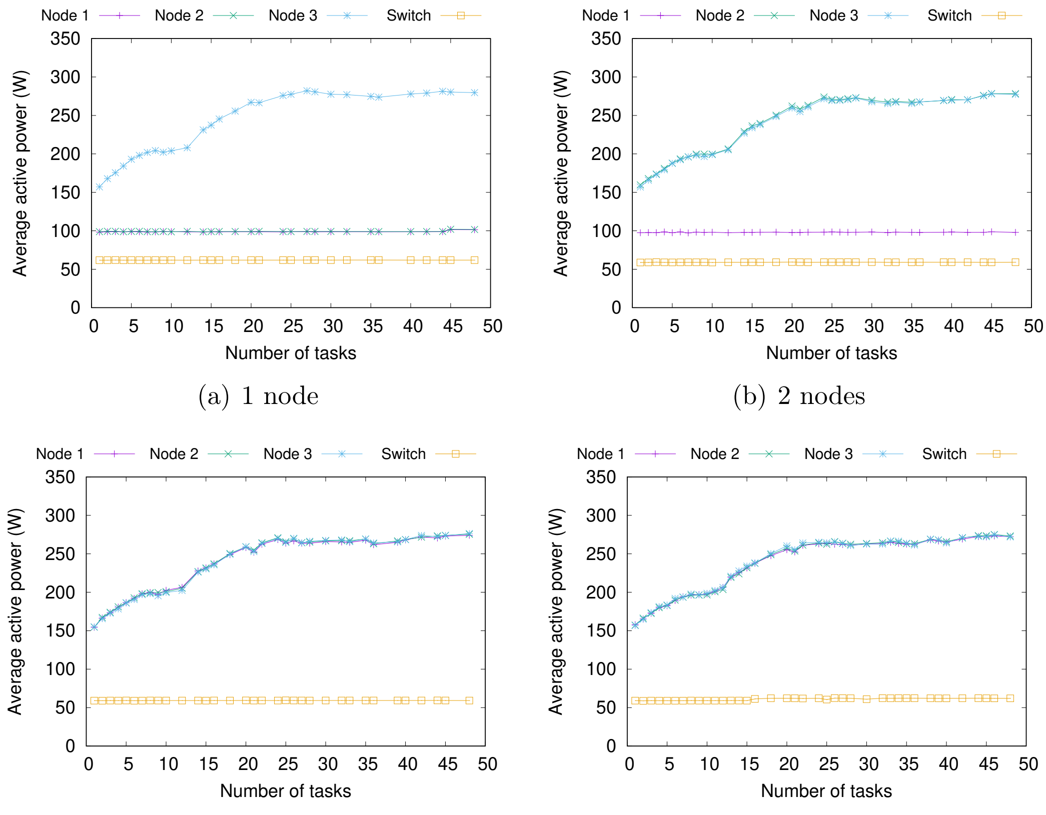
Figure 2: Node power consumption for application Graph500 varying the number of tasks



(a) 1 node (b) 2 nodes

(c) 3 nodes (d) 6 nodes

Figure 3: Node power consumption for application Linpack varying the number of tasks



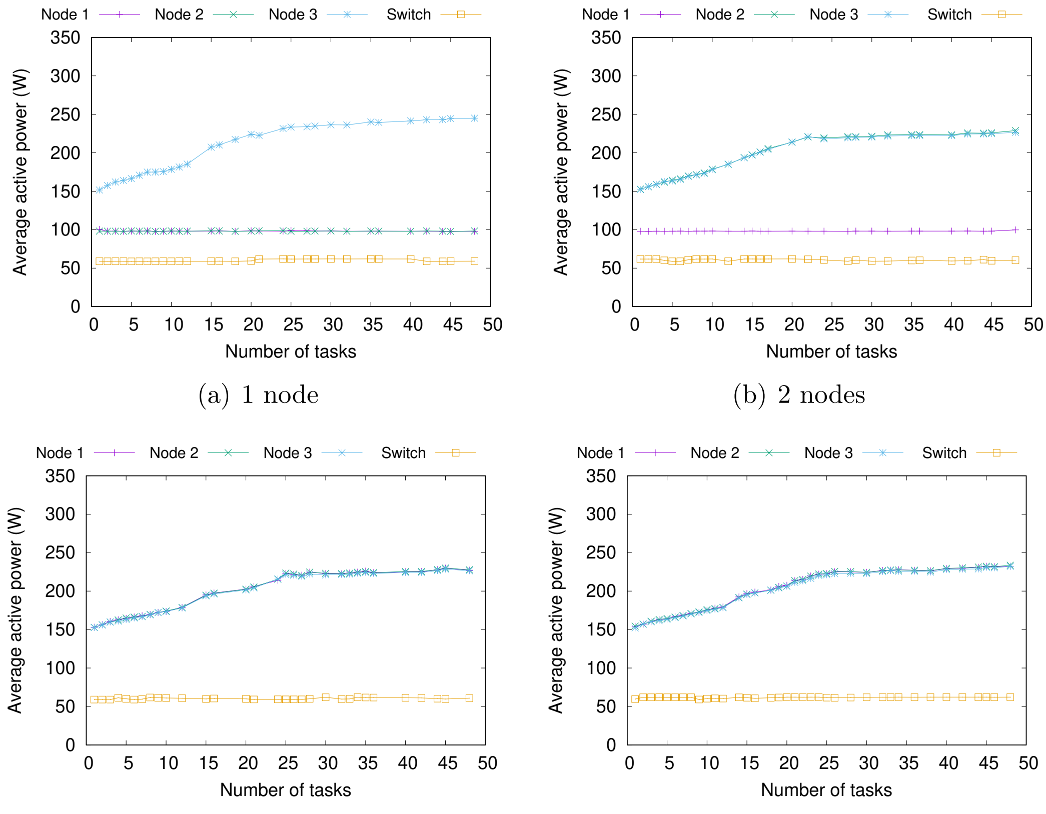
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 4: Node power consumption for application HPCG varying the number of tasks



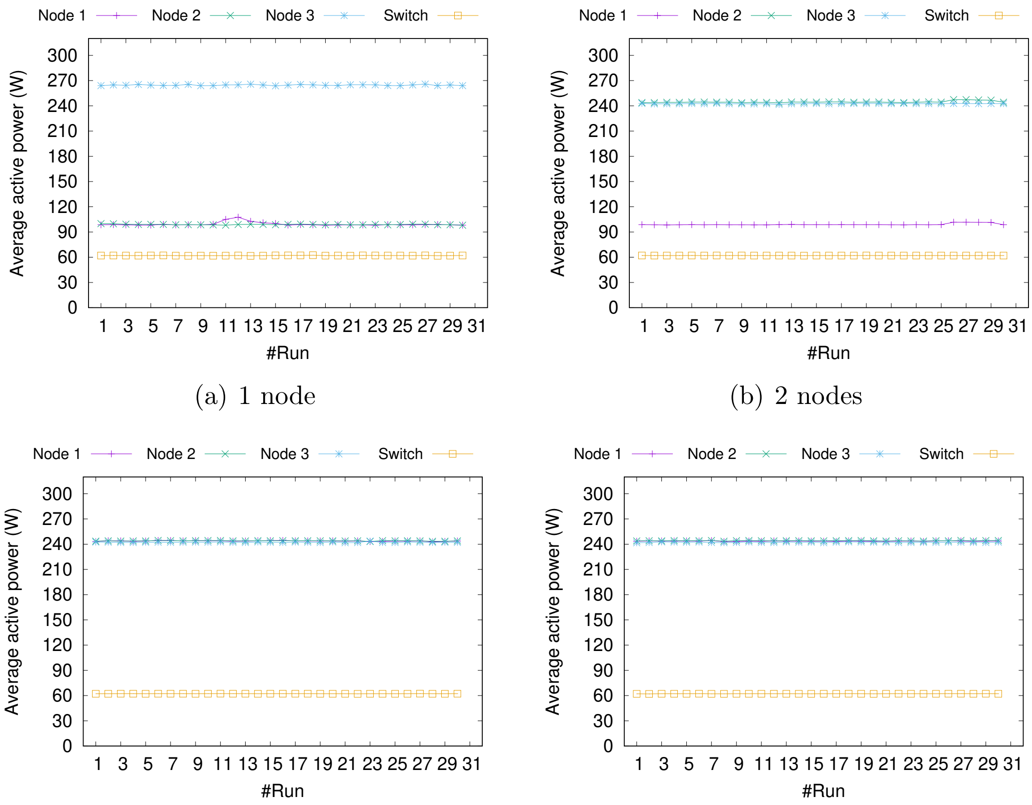
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 5: Node power consumption for application Gromacs varying the number of tasks



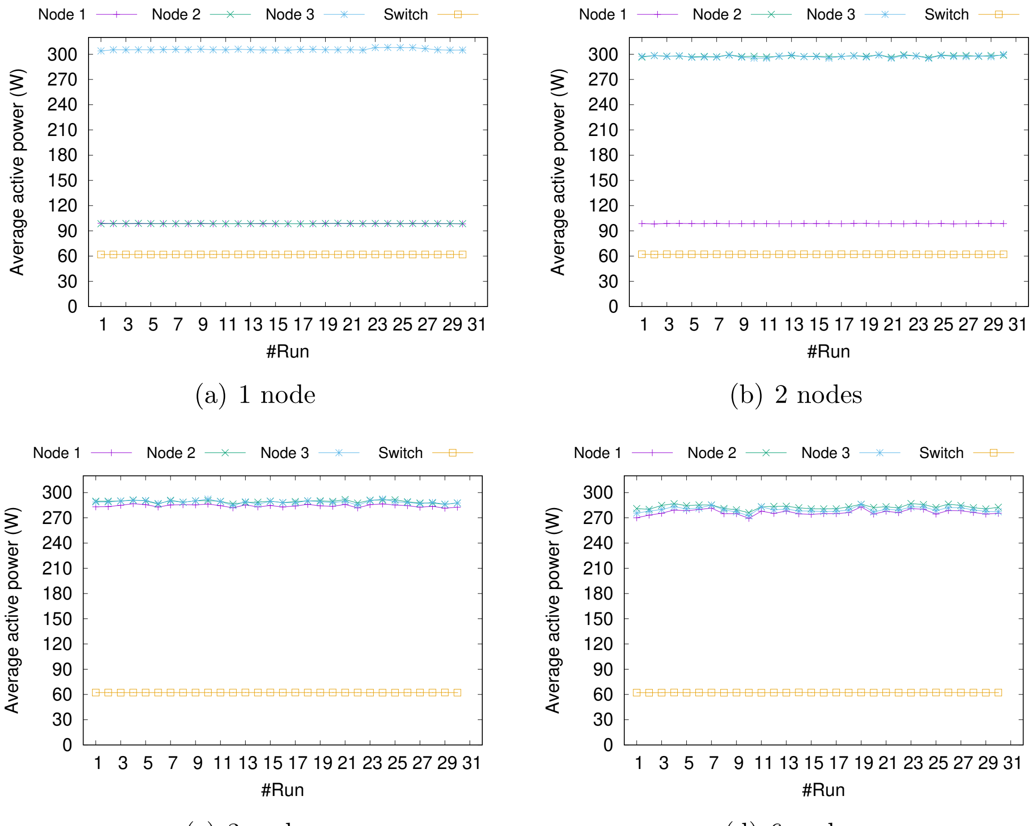
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 6: Node power consumption for application Graph500 with 48 tasks per node



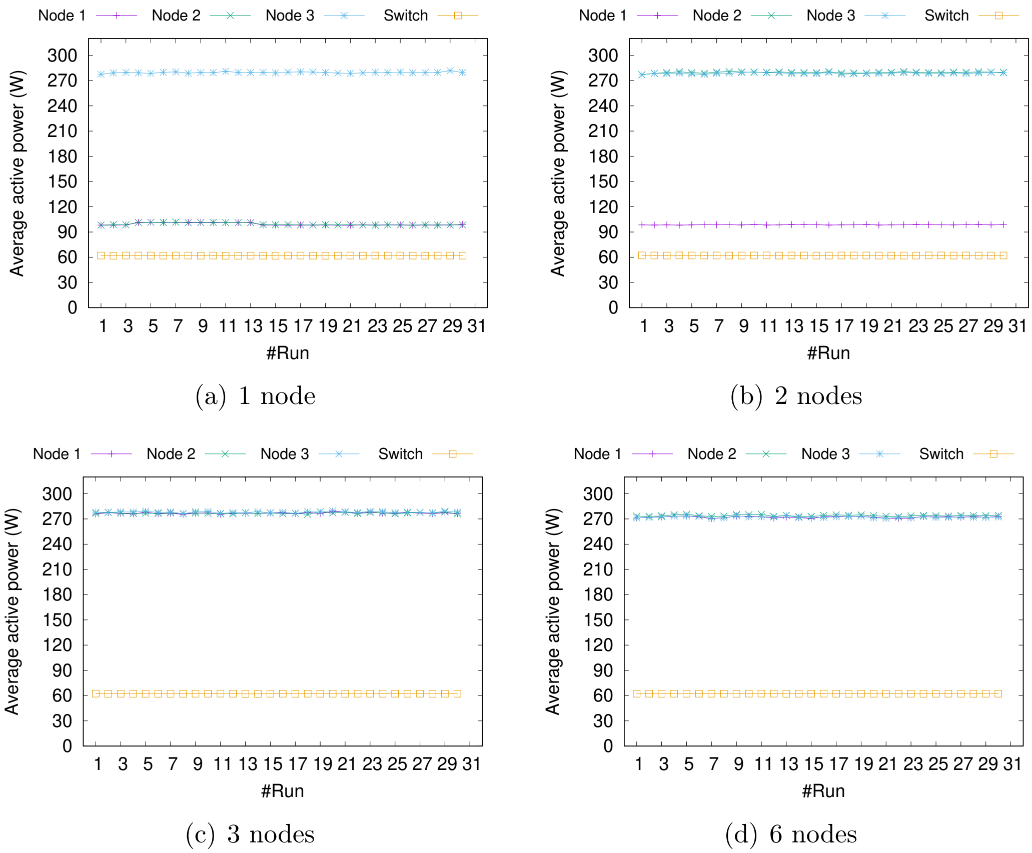
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 7: Node power consumption for application Linpack with 48 tasks per node



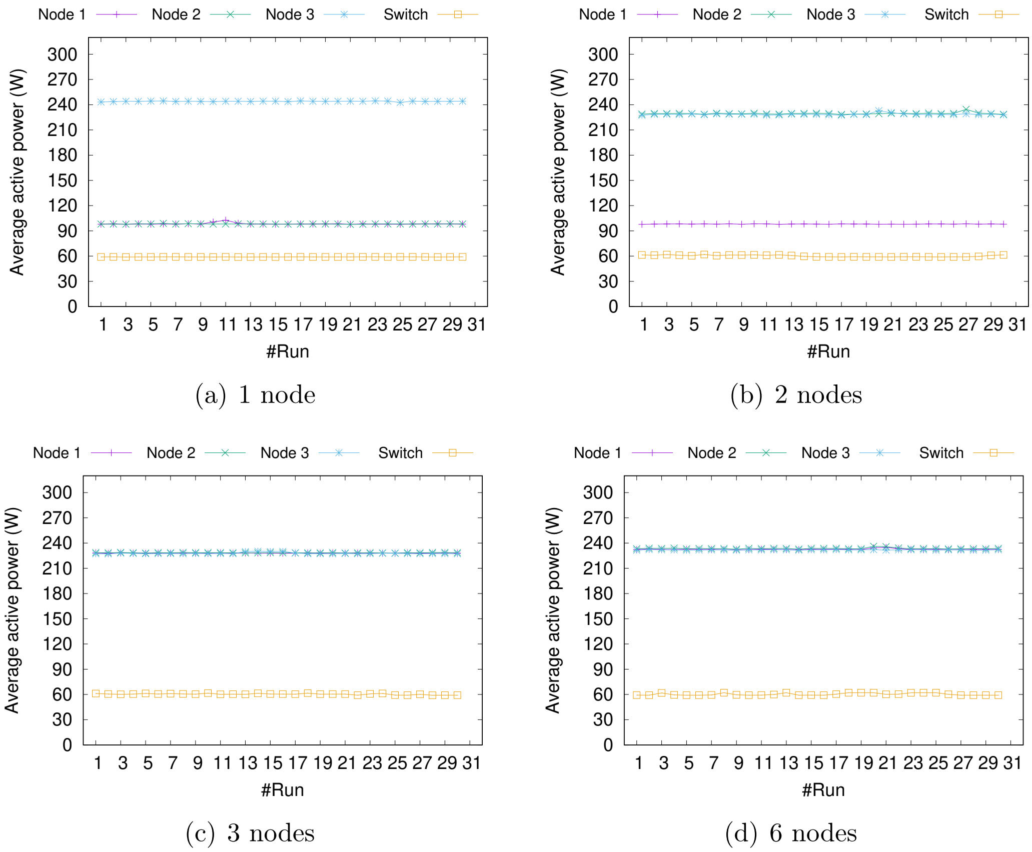
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 8: Node power consumption for application HPCG with 48 tasks per node



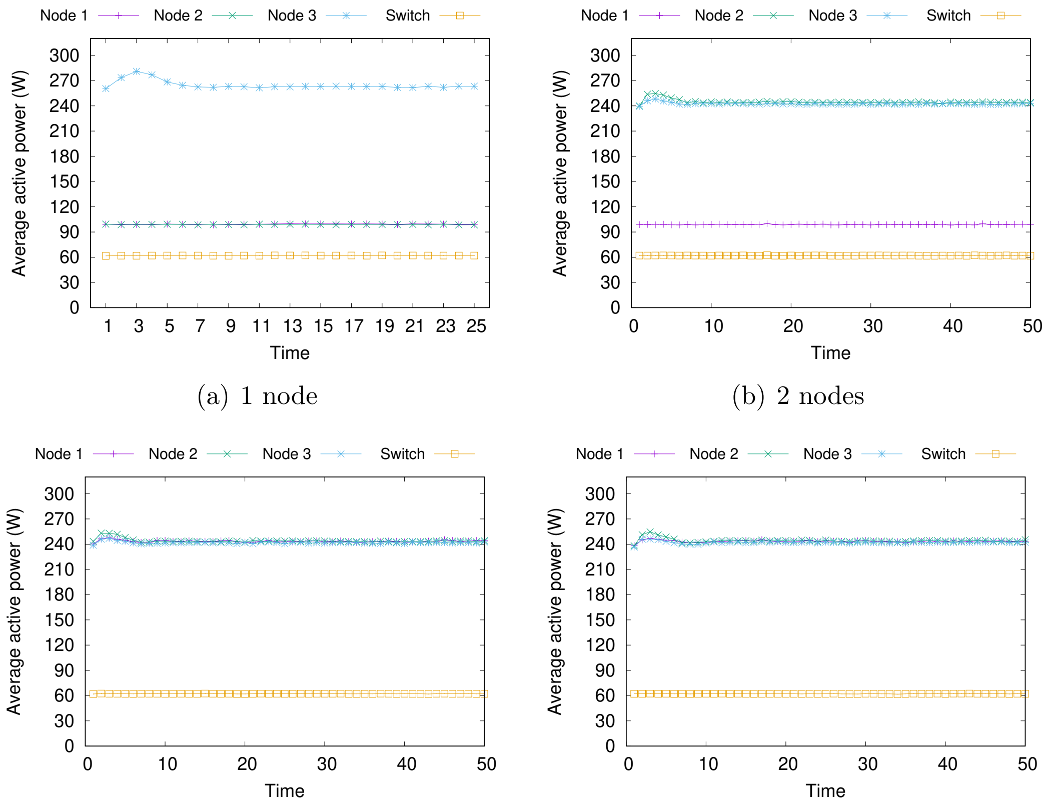
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 9: Node power consumption for application Gromacs with 48 tasks per node



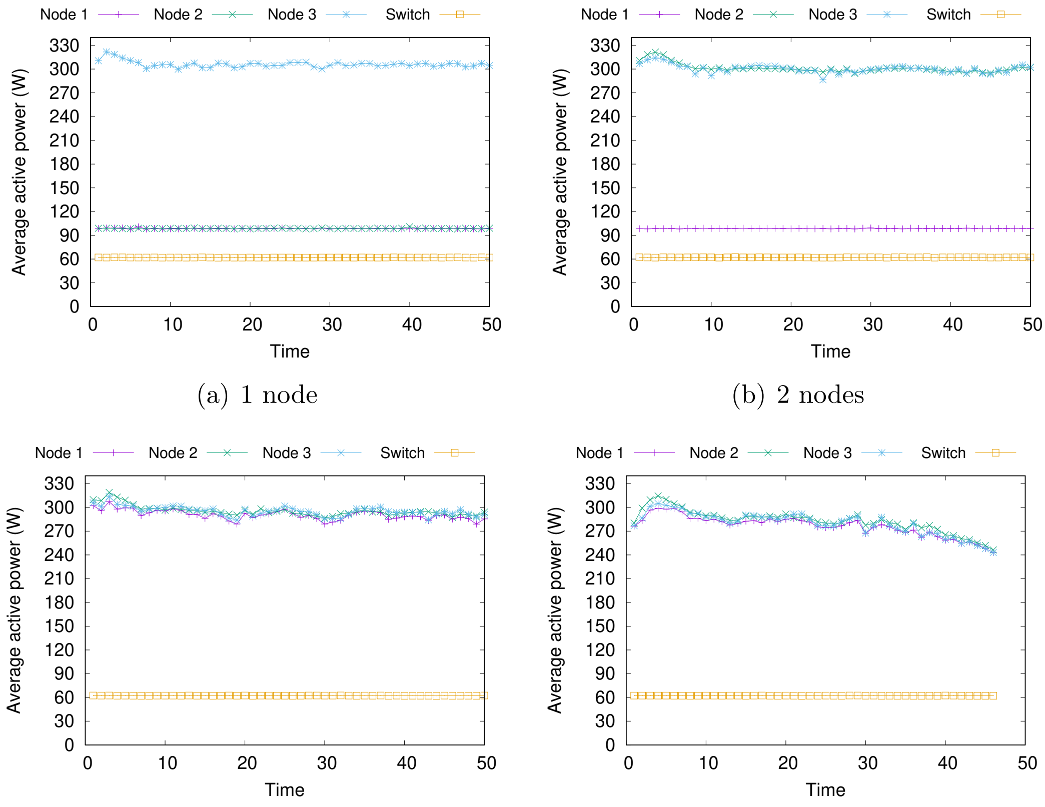
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 10: Node power consumption for application Graph500 over time



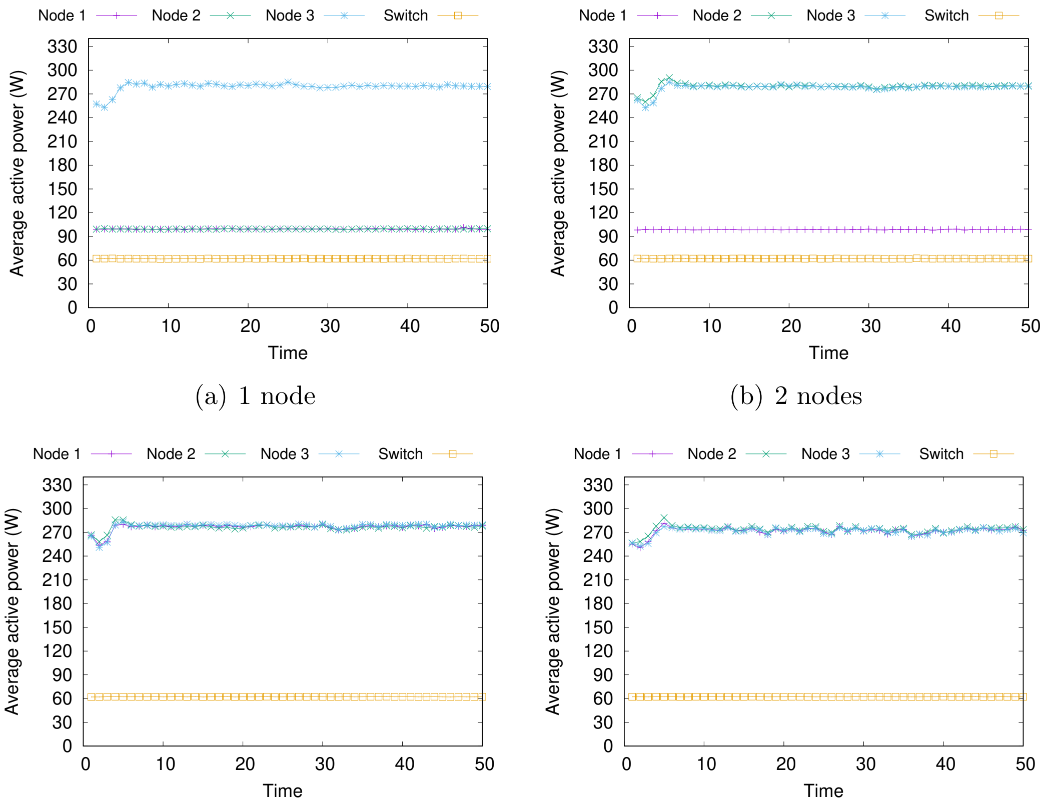
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 11: Node power consumption for application Linpack over time



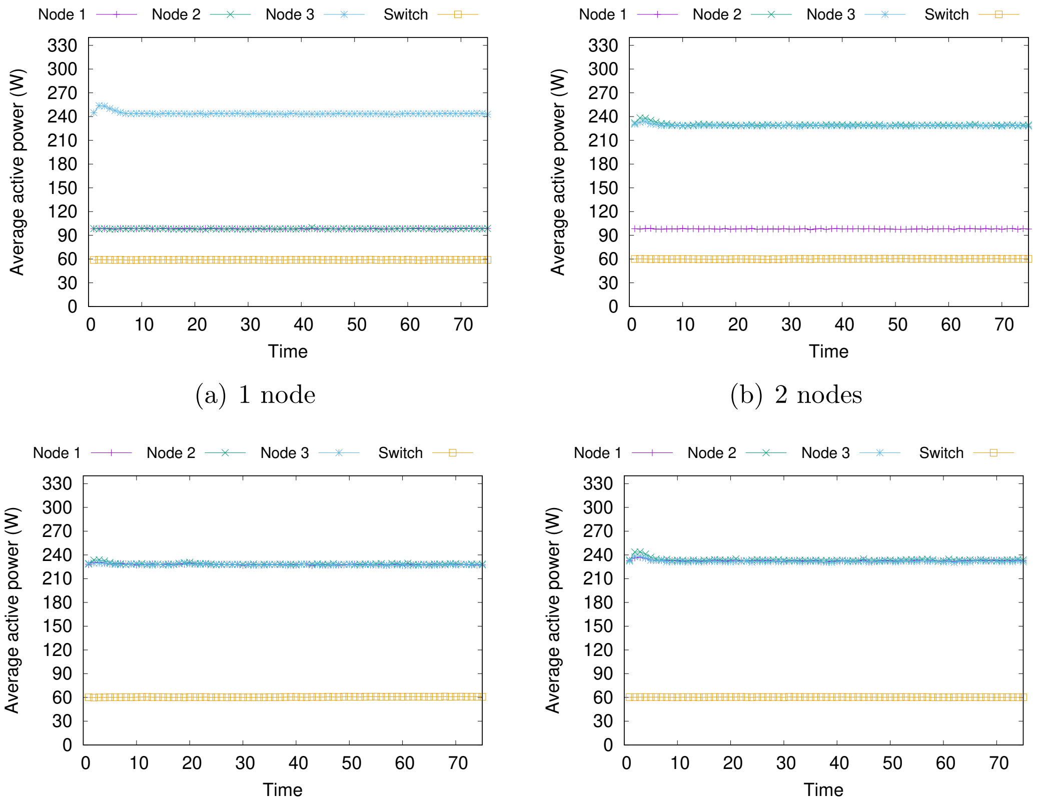
(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

Figure 12: Node power consumption for application HPCG over time



(a) 1 node

(b) 2 nodes

(c) 3 nodes

(d) 6 nodes

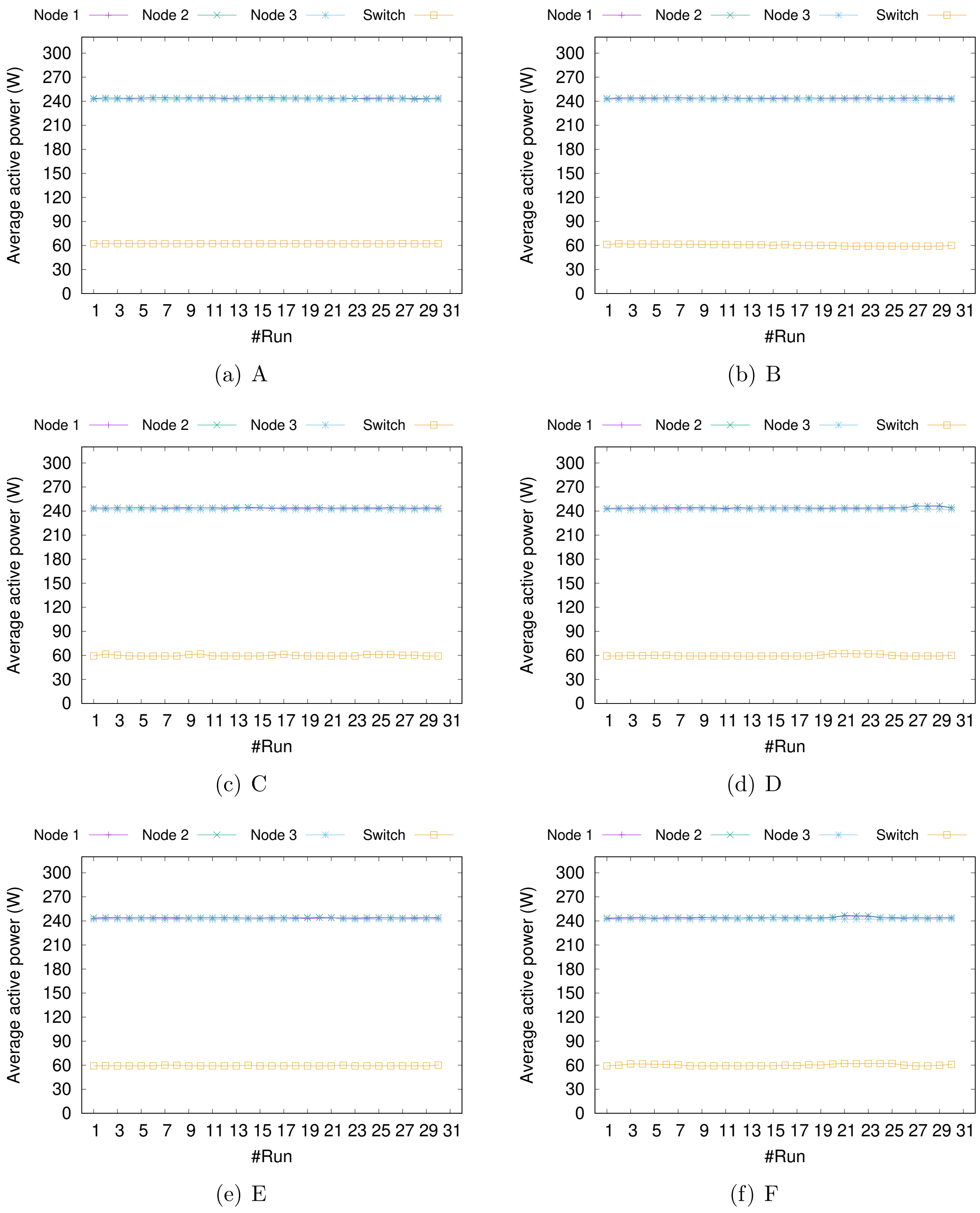
Figure 13: Node power consumption for application Gromacs over time

Table 2: Task assignments.

	Node 1	Node 2	Node 3
A	$0 \cdots 23$	$24 \cdots 47$	$48 \cdots 71$
B	$0 \cdots 23$	$48 \cdots 71$	$24 \cdots 47$
C	$24 \cdots 47$	$0 \cdots 234$	$48 \cdots 71$
D	$24 \cdots 47$	$48 \cdots 71$	$0 \cdots 23$
E	$48 \cdots 71$	$0 \cdots 23$	$24 \cdots 47$
F	$48 \cdots 71$	$24 \cdots 47$	$0 \cdots 23$

As indicated above, the PZ4000 power analyzer has only 4 modules to collect data of power consumption. In our experiment, one of this modules is always dedicated to the switch, and therefore only three of the six available nodes can be monitored. Note that the six available nodes are totally equal in their configuration. However, as we are interested in checking if there are differences in node power consumption along a run, and it is not operative to change the meter wiring, we have instead chosen to vary the task mapping to nodes, in order to be able to measure the power consumption when other tasks are executed in each node.

Figures 14, 15, 16 and 17 show the results obtained for all the applications. Each application runs considering 144 tasks in total, 48 in each of the three nodes considered in the test. In each one of them the task assignment is different. As can be seen, the results in all cases allow us claim that the power consumption in each node is the same, regardless of the tasks it is in charge of.



(a) A

(b) B

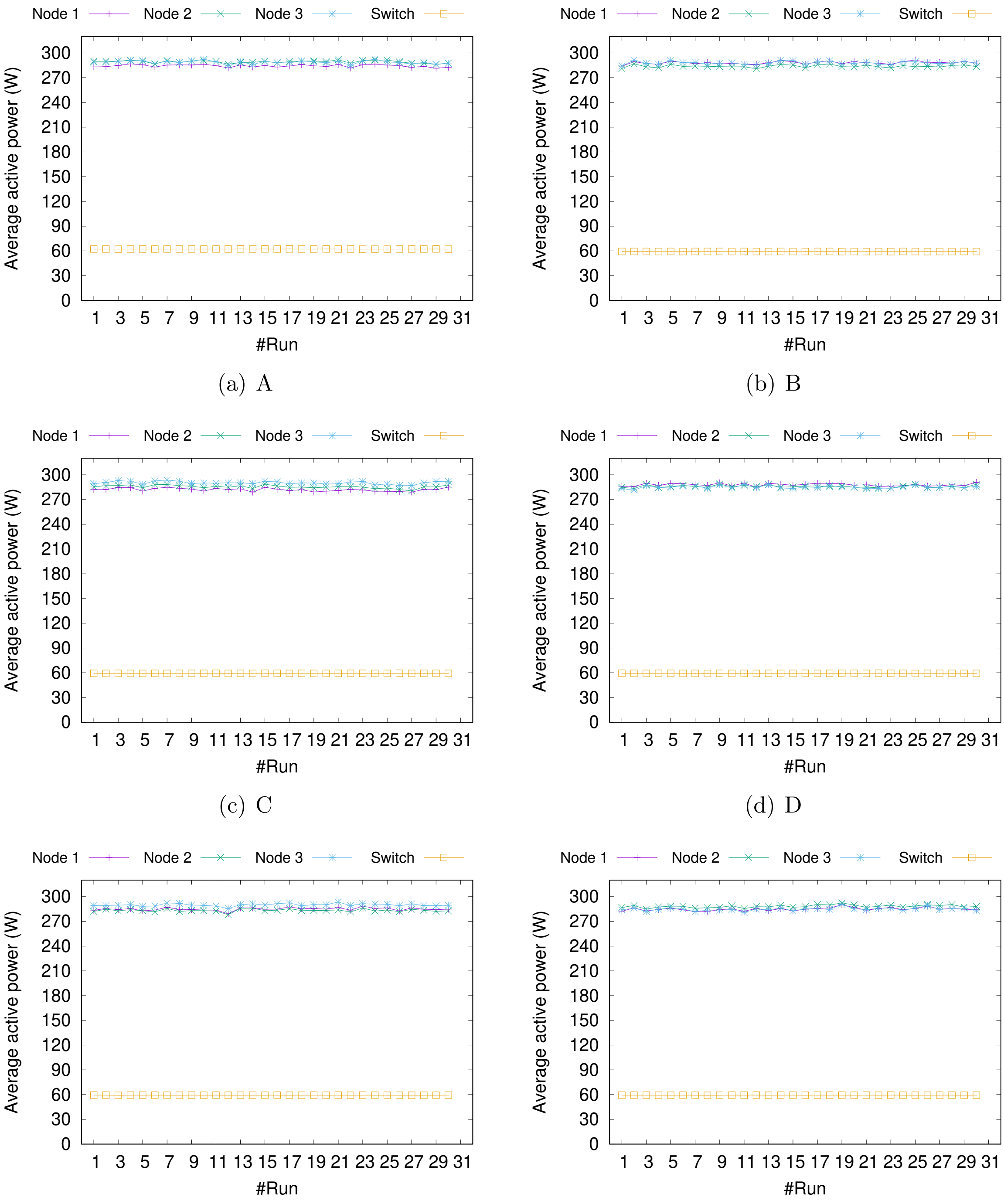
(c) C

(d) D

(e) E

(f) F

Figure 14: Node power consumption for application Graph500 when varying task mapping



(a) A

(b) B

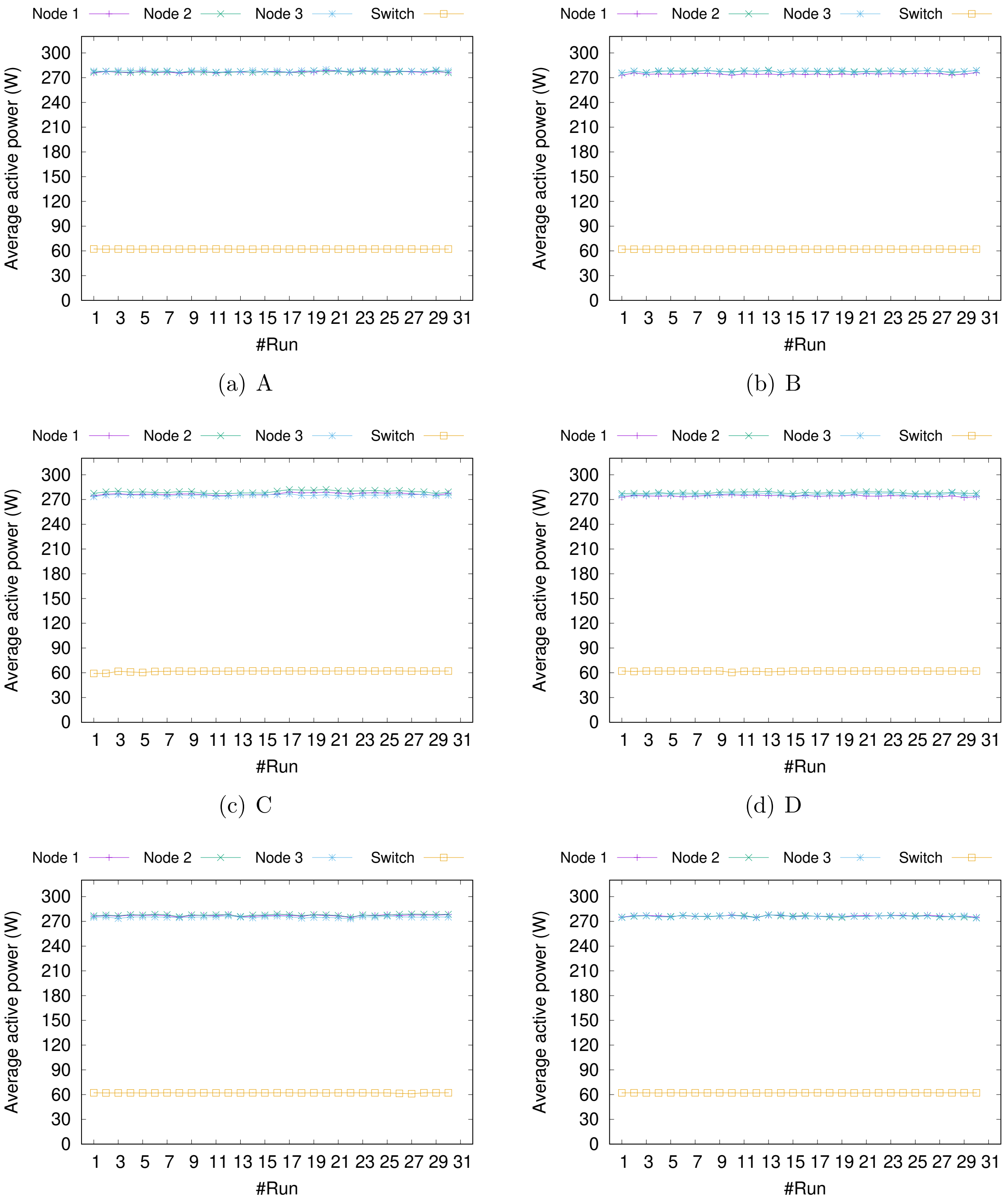
(c) C

(d) D

(e) E

(f) F

Figure 15: Node power consumption for application LinPack when varying task mapping



(a) A

(b) B

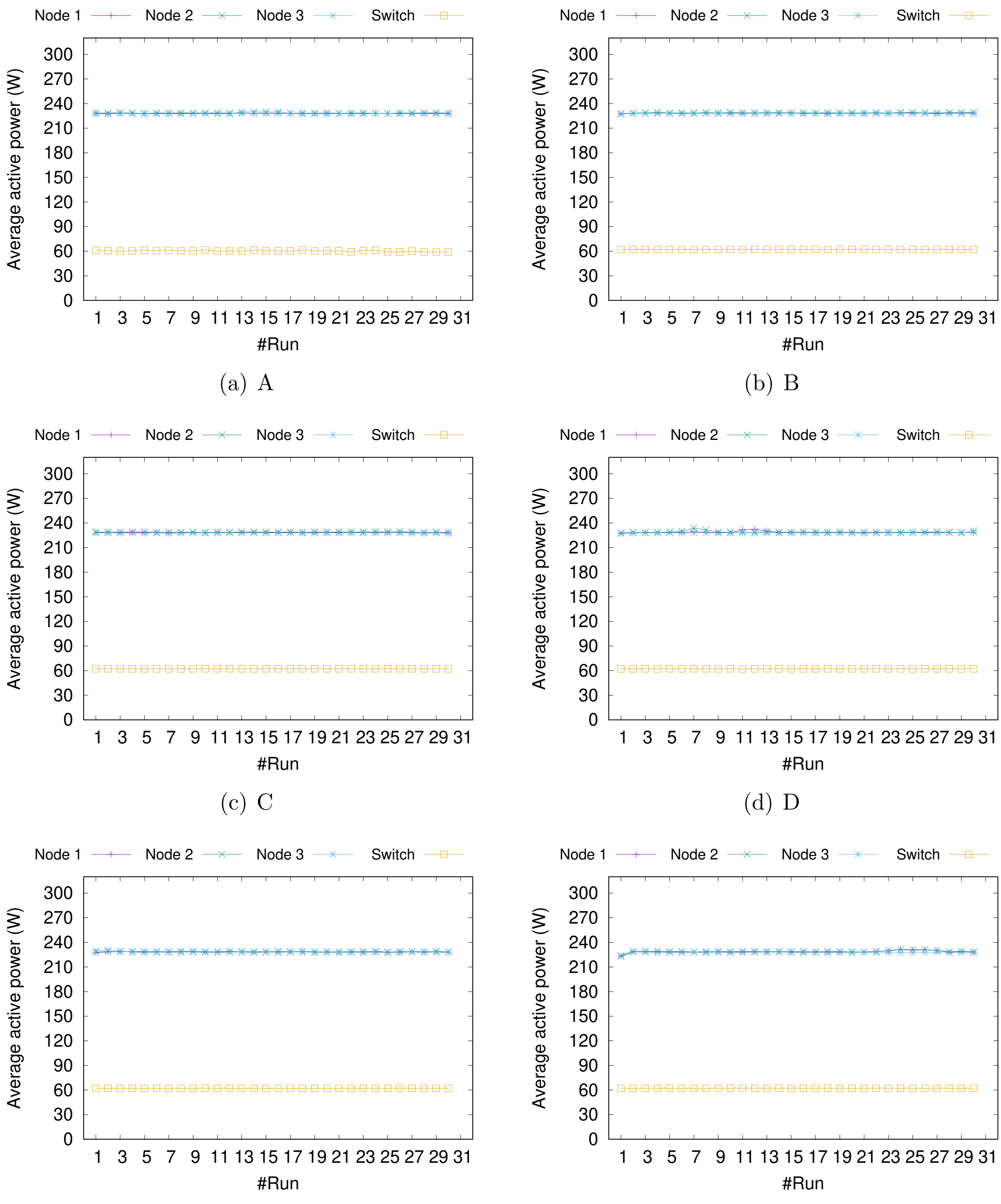
(c) C

(d) D

(e) E

(f) F

Figure 16: Node power consumption for application HPCG when varying task mapping



(a) A

(b) B

(c) C

(d) D

(e) E

(f) F

Figure 17: Node power consumption for application Gromacs when varying task mapping

4.2 Switch power consumption

To obtain the power consumption of the switch, measurements have been made by varying the number of ports used. Initially there are no nodes connected to the switch but latter then nodes are connected one by one, up to 36. Figure 18 shows the data, both graphically and numerically.

As can be seen, on the one hand, there is a fixed power consumption even if no node is connected to the switch (number of ports 0 in the figure). On the other hand, the power consumption increases linearly with the number of ports used. Specifically, it can be seen that each port contributes approximately in 4 watts to the total consumption of the switch. The figure shows the average values and the confidence intervals at 95% confidence level of all power consumption data collected for the nodes and the switch of the test bed configuration. As you can see, the variation in data for each component is minimal, and the confidence interval is very narrow.

This test has been performed in two different ways. Firstly, the end of the cable connecting the node to the switch has been disconnected/connected from the switch. In a second case, the disconnection/connection has been done at the end of the node. In this second case, the port contribution to the switch's power consumption is the same as if the node were connected to the cable. Results in Figure 18 have been obtained in this way.

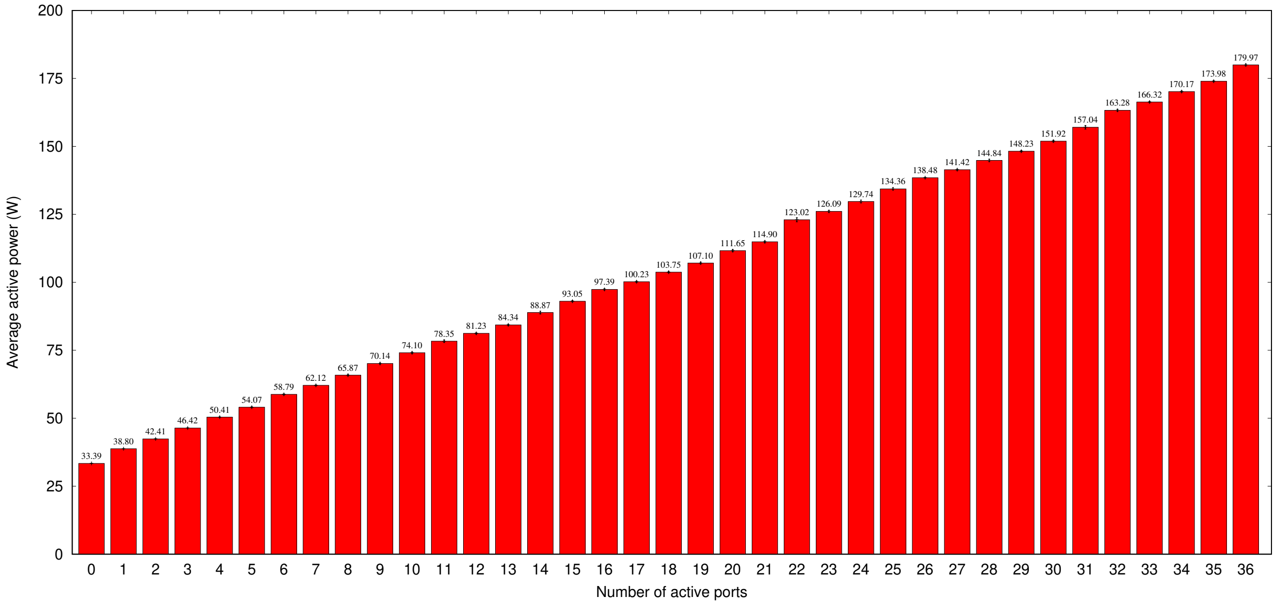


Figure 18: Switch power consumption.

4.3 Network interface power consumption

In order to determine the power consumption of a network interface (NIC), a couple of tests have been carried out. Firstly, without launching any application, the power consumption in the nodes is measured, and then the test is repeated by removing the node NIC. Secondly, same tests as before, but launching an application on a node.

The results are shown in Figures 19 and 20, respectively. It can be seen that when the node does not have NIC, its consumption is reduced by approximately 10 watts.

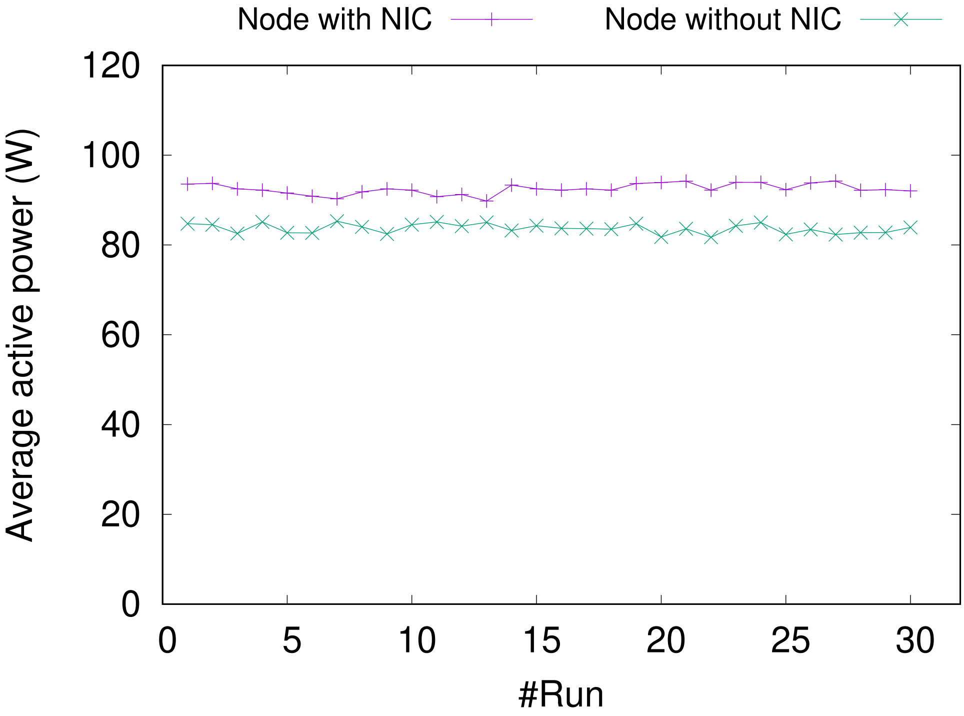


Figure 19: Power consumption of the network interface when the node is idle.

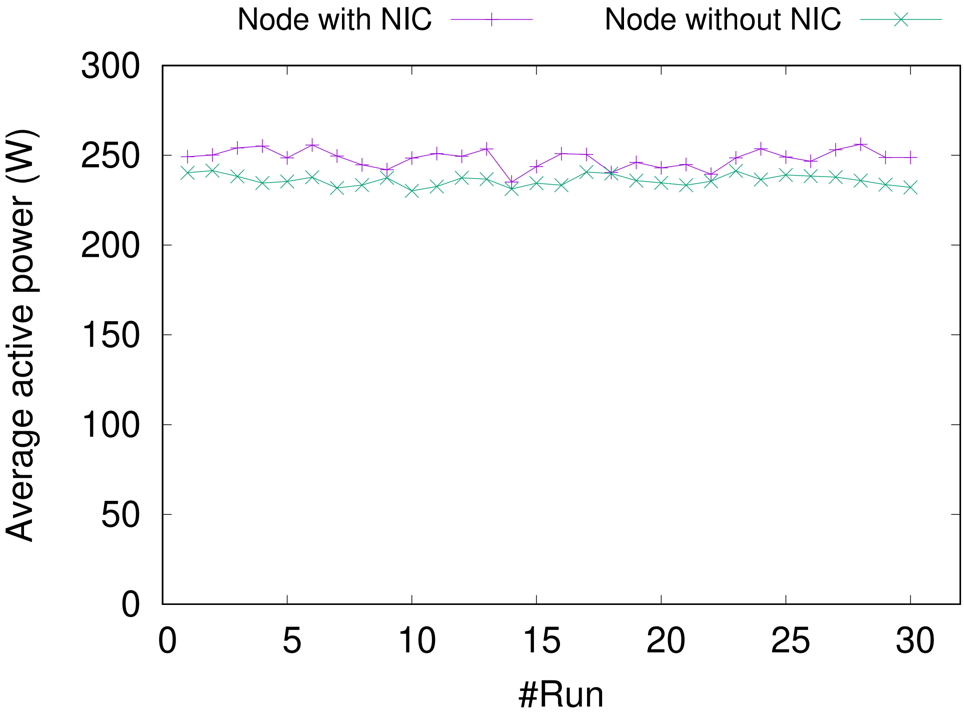


Figure 20: Power consumption of the network interface when the node is running an application.

4.4 Resume

Figure 21 shows a summary of the power consumption data collected in this study. For each application, the following three results are indicated: the average switch consumption, the maximum node power consumption (red bar) when the applications run only on one node (multiple processes), and the maximum node power consumption (green bar) when several nodes (the number is indifferent) are involved in the execution of the applications.

Figure 21 shows the average values and the confidence intervals at 95% confidence level of all power consumption data collected (30 runs per each). As you can see, the variation in data for each component is minimal, and the confidence interval is very narrow.

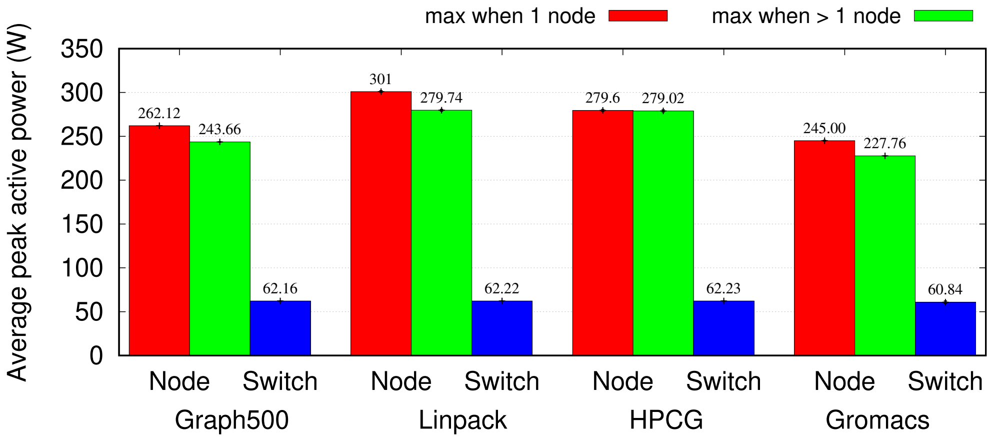


Figure 21: Average peak power consumption of the node and switch.

5 Conclusions

The power consumption of HPC systems represents a large fraction of the cost of maintenance of this kind of infrastructures. It is very important to reduce the total cost, and therefore new proposals for that are required. These proposals are generally tested and evaluated through simulation by using power consumption models, which need power consumption data from the different components of real HPC systems.

We have collected power consumption data for the main elements of the computing cluster part, specifically compute nodes, switch and network interface. For that, we have used dedicated external power measurement hardware. In this report we show all the data collected, which have been obtained from the execution of various HPC applications.

These are some of the most important conclusions that can be drawn from the data:

- The compute nodes are the components that consume the most. Although the power consumption of the network interface card is included in the consumption of the nodes when the data have been shown, its contribution (between 5 and 10 watts) is very small compared to the total power consumption of the node.
- The nodes have a fixed base power consumption, which is approximately 100 watts for the switch considered in this study. This is the case when there are no applications running on the nodes.

- Power consumption increases linearly with the utilisation of its resources. In particular, this linear increase is clearly observed when the number of cores used to run the application is progressively increased.
- The switch has a fixed power consumption, even if no node is connected to the switch.
- The total power consumption of a switch is directly proportional to the number of ports it has, there being a linear relationship between both. For the switch used, it has been verified that a port contribution to the switch's power consumption occurs whether the cable is connected to the node or not.

References

[1] The Graph500 list. `https://graph500.org`, November 2017.

[2] L. A. Barroso and U. Hölzle. The case for energy-proportional computing. *Computer*, 40(12):33–37, Dec 2007.

[3] HPC challenge benchmark. `http://icl.cs.utk.edu/hpcc/index.html`.

[4] K. Masahiro, T. Hirotaka, K. Katsuhiro, and K. Kazuo. PZ4000 power analyzer. Technical report, 2001.

[5] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[6] A. Shehabi, S. J. Smith, D. A. Sartor, R. E. Brown, M. Herrlin, J. G. Koomey, E. R. Masanet, N. Horner, I. Lima Azevedo, and W. Lintner. United states data center energy usage report. Technical report, Jun 2016.

[7] Yokogawa Electric Corporation. *PZ4000 Power Analyzer. User's manual*, 2002.