An Analysis of Energy-Aware in HPC Environments including Grid'5000 in France

Authors: Laércio Pioli Jr., Eduardo C. Inácio, Mario Antônio R. Dantas, Victor Ströele A. Menezes

Institutes: Knowledge Engineering Research Center- NEnC at Federal University of Juiz de Fora (UFJF), and Research Laboratory in Distributed Systems - LAPESD at Federal University of Santa Catarina (UFSC)

Postgraduate program in Computer Science
www.pgcc.ufjf.br

Contents

- Introduction

- How to measure the energy consumption in HPC Environment?
- According to Kamil et al. [4], there are many possible ways to obtain power usage.

- Line Meters.
- Clamp Meters.
- Power Panels.
- Software Measurements.

- GREEN500 provide a ranking of the most energy-efficient super-computers in the world.

- Introduction

Green500 List uses “performance per watt” (PPW) as its metric to rank the energy efficiency of supercomputers. [2]

$$PPW = \frac{\text{Performance}}{\text{Power}}$$

Figure: Performance per Watt Metric

- Introduction

- Six of the eight Green500 lists had RIKEN supercomputers as the most energy-efficient system since June-2015

Green500 Rank	MFLOPS/W	Site	System	Total Power(kW)
1	7031.4	RIKEN	ExaScaler-1.4 80Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SC	50.3
2	6841.3	High Energy Accelerator Research Organization /KEK	ExaScaler-1.4 16Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband, PEZY-SC	28.3
3	6217.9	High Energy Accelerator Research Organization /KEK	ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC	32.6
4	5272.1	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.2
5	4258.1	GSIC Center, Tokyo Institute of Technology	LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	39.8

Figure: Green500 List for June 2015 [1]

[1] https://www.top500.org/green500/lists/2015/06/

- Introduction

Green500 Rank	MFLOPS/W	Site	System	Total Power(kW)
1	7031.4	Institute of Physical and Chemical Research (RIKEN)	ExaScaler-1.4 80Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SC	50.3
2	5331.5	GSIC Center, Tokyo Institute of Technology	LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.1GHz, Infiniband FDR, NVIDIA Tesla K80	51.1
3	5272.1	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.2
4	4778.5	Institute of Modern Physics (IMP), Chinese Academy of Sciences	Sugon Cluster W780I, Xeon E5-2640v3 8C 2.6GHz, Infiniband QDR, NVIDIA Tesla K80	65
5	4112.1	Stanford Research Computing Center	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190

Figure: Green500 List for November 2015 [2]

[2] https://www.top500.org/green500/lists/2015/11/

- Introduction

Green500 Rank	MFLOPS/W	Site	System	Total Power(kW)
1	6673.8	Advanced Center for Computing and Communication, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	150.0
2	6195.2	Computational Astrophysics Laboratory, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	46.9
3	6051.3	National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	15371
4	5272.1	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.2
5	4778.5	Institute of Modern Physics (IMP), Chinese Academy of Sciences	Sugon Cluster W780I, Xeon E5-2640v3 8C 2.6GHz, Infiniband QDR, NVIDIA Tesla K80	65

Figure: Green500 List for June 2016 [3]

[3] https://www.top500.org/green500/lists/2016/06/

- Introduction

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	259	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	842.0	50	17.009
2	307	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	788.2	47	16.759
3	276	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	149	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	1,350	14.173

Figure: Green500 List for November 2017 [4]

[4] https://www.top500.org/green500/lists/2017/11/

- Introduction

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	359	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	857.6	47	18.404
2	419	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	798.0	47	16.835
3	385	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	227	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	1,649	14.423

Figure: Green500 List for June 2018 [5]

[5] https://www.top500.org/green500/lists/2018/06/

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	375	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	953,280	1,063.3	60	17.604
2	374	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
3	1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,397,824	143,500.0	9,783	14.668
4	7	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	1,649	14.423
5	22	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	135,828	8,125.0	792	13.704

Figure: Green500 List for November 2018 [6]

[6] https://www.top500.org/green500/lists/2018/11/

- Introduction

The world's greenest supercomputer today.

The Shoubu system B., located at RIKEN, has appeared tree times consecutively in the Green500 list as the most energy efficient system. (NOV-2017, JUNE-2018 and NOV-2018) [a]

It's a ZettaScaler-2.2 supercomputer installed at the Advanced Center for Computing and Communication, RIKEN, Japan.

---

[a] https://www.top500.org/green500/

Contents

- Goal

We are interested in analyzing the power consumption of the I/O requests through different approaches and configurations including:

- Network interconnects
- Cluster Location
- Differents Workloads Scenarios
- Storage and Network Devices
- Distributed Parallel File Systems.

Contents

High Performance Computer (HPC) environments, usually uses Distributed File System (DFS) as the structure.

As we know, DFS is a kind of file system. According to Silverschatz et al. [7], a file system "consists of two distinct parts: a collection of files, each storing related data, and a directory structure, which organizes and provides information about all the files in the system."

A file is a named collection of related information that is recorded on secondary storage.

In general, a file is a sequence of bits, bytes, lines, or records, the meaning of which is defined by the file's creator and user.

The directory structure also has the function of organizing the files in the system.:

File systems provide efficient and convenient access to the disk by allowing data to be stored, located, and retrieved easily.

- Distributed File Systems -layers

According to Silverschatz et al. [7], the file-system implementation consists of three major layers, as depicted schematically bellow.

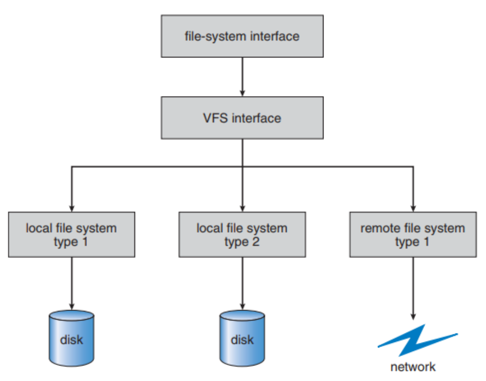


Figure: Schematic view of a virtual file system

DFS can share files between process which may even be running on different computers. Some DFS architectures are shown below:

- Client-Server Architecture.
- Cluster-Based Architecture (Parallel File System).
- Symmetric Architecture.

- Distributed File Systems -xxxxxxx

- Client-Server Architecture.

  - Network File System (NFSv3) [1] is one of the most widely-deployed DFS for UNIX-based systems.[8].
  - The basic idea behind NFS is that each file server provides a standardized view of its local file system.

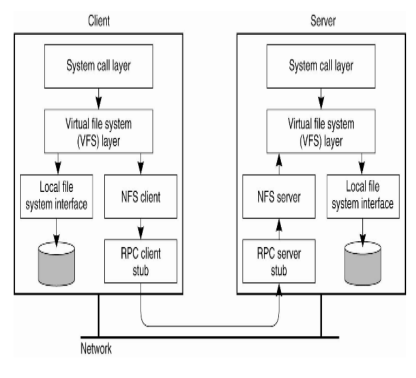


Figure: NFS architecture for UNIX modern systems

- Distributed File Systems -

- Cluster-Based Architectur also known as (Parallel File System).

- Server clusters are often used for parallel applications.

- By distributing a file across multiple servers, it becomes possible to process different parts in parallel.

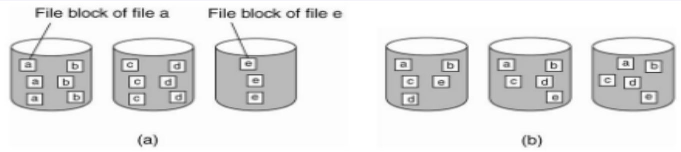


Figure: (a) Distributing whole files across several servers and Striping files for parallel access.

- Distributed File Systems -xxxxxxx

- Symmetric Architecture.

- The peer-to-peer file system Ivy [5] does not have a central element that is responsible for managing the file system.[8].
- Their system essentially consists of three separate layers as shown.
-Implementations through Hash Tables.

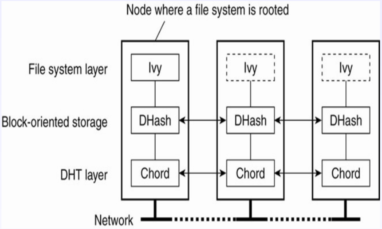


Figure: Ivy distributed file system

- Distributed File Systems -Components of a PFS

- PFS are composed basically by 3 requirements: data server, meta-data server and client. Shan et al. [6]

- Data server is the component responsible for the persistence of the contents of the files.

- Metadata server keeps information about the files (metadata) updated.

- Client is the component that enables interaction with the parallel file system.

Contents

Power Measurement Methods - How can we energy measure?

According to Kamil et al. [4], there are many possible ways to obtain power usage.

- Line Meters.
  Hardware which is connected to the line. The power whips are bolted down on both ends.
  - Could measure precisaly the power usage.
  - Require disconnecting the system to be measured.
- Clamp Meters.
  Electric device with jaws that open to allow attachment around an electric conductor.
  - Provide a way to measure power without needing to disconnect a system.
  - Can only be used to measure current on individual conductors of a 2-phase or 3-phase multi-conductor cable.

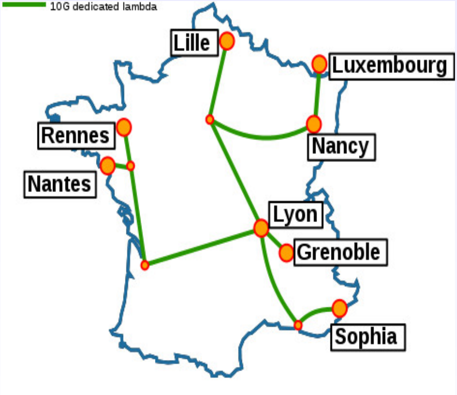
Power Measurement Methods - How can we energy measure?

- Power Panels.
  Native hardware that is coupled to the unit.
  - Provide the only way to monitor large pieces of an HPC system or the entire HPC system itself.
  - Not provide finegrained measurements of individual pieces of a system.
- Software Measurements.
  Allows for greater granularity of measurement and does not require physical access.
  - Can exploit much as possible some specific hardware energy consumption (CPU, Memory, Network, IO Subsystem).
  - Not accurately as physical devices.

Contents

- Experimental Testbeds

- Where are the experiments are being runned?



"Grid5000 is a large-scale and versatile testbed for experiment-driven research in all areas of computer science, with a focus on parallel and distributed computing including Cloud, HPC and Big Data."

Figure: Grid5000[a]

[a] https://www.grid5000.fr

- Experimental Testbeds - Grid5000 Key Features

Provides access to a large amount of resources:[a]

- 8 sites - Grenoble, Lille, Luxembourg, Lyon, Nancy, Nantes, Rennes, Sophia-Antipolis.
- 33 clusters
- 1064 compute-nodes grouped in homogeneous clusters.
- 12952 CPU cores
- 88 GPUs
- 208 SSDs and 1188 HDDs on nodes (total: 1126.93 TB)

---

[a] https://www.grid5000.fr/mediawiki/index.php/Hardware

Contents

- Assessment Considerations. -Network Hardware

- Networks Interconnects:

- The computer-networking interconnect provided by Grid'5000 sites are:
- Ethernet (1, 10 and 25 Gbps each)
- InfiniBand (20, 40 and 56 Gbps each)
- Omni-Path (100 Gbps)

- For instance, Grenoble site has 2 clusters, Yety and Dahu, of which are locately at the same place and are using both 10 Gbps + 100 Gbps Omni-Path Network.

- In contrast, Nancy site has graphene and grele clusters, which has 1 Gbps + 20 Gbps InfiniBand and 10 Gbps + 100 Gbps Omni-Path respectively.

- Clusters Location:

- We are considering clusters and nodes which are located in the same geographic place as well in different geographic place.

- For exemple, we can consider the energy consumption to execute workloads through clients and servers which are located in the same cluster, or even located in clusters geographically distant.

- We can analyze the energy consumption between these clusters and even with a greater variation of clusters.

- Storage Device:

- We are considering different storage devices to perform our experiment.

- We can use and distribute these devices through the organization of the storage architecturein in the file systems components.

- Currently used devices.
- HDD's.
- SSD's.

- I/O Workloads Generator.

- IOR - ITERLEAVED-OR-RANDOM (IOR)[6]
A benchmark designed to replicate regular access patterns by generating I/O requests from the same size to contiguous data blocks where the size must be multiple of the request size.

- IORE - IOR-EXTENDED [3]
Parallel I/O workload generator that reproduces irregular and consequently more complex access patterns, where the sets of processes can be of heterogeneous loads unlike the IOR.

- Application I/O Kernels.

- BT-IO - Block-Tridiagonal, [9] Is a pseudo application which is part of a small set of programs, more specifically a benchmark, called NAS Parallel Benchmarks (NPB). It is used to perform the output capability of high performance computing systems, specifically parallel systems [9].

- VPIC - Vector Particle-In-Cell, is a fully relativistic plasma simulation code which is provided by Los Alamos National Laboratory (LANL) and is operated by Los Alamos National Security, LLC for the U.S. Department of Energy. The source code and aditional information can be accessed in github through the link below. [a].

[a] https://github.com/vpic/vpic

- Power API provided by the Grid'5000 cluters.

- PowerAPI is a middleware toolkit for building software-defined power meters [a]. Some nodes has a Power Distribution Units (PDU), which is a device that suplyies the eletrical power.

- Kwapi is a tool that provides a convenient and consistent way to monitor energy consumption in experiments.

[a] http://powerapi.org/

Parallel File Systems selected for the experiment:

- Orange File System.
- Lustre File System.
- Ceph File System.

Contents

- Acknowledgment

- RIKEN - Japan
  We would like to thanks the RIKEN Institute for supporting this presentation and to give us the oportunity to shown our work. [a]
- GRID'5000 - France
  The experiments presented in this study were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.[b]

[a] http://accc.riken.jp/en/

[b] https://www.grid5000.fr

References I

Brent Callaghan. *NFS illustrated*. Addison-Wesley, 1999.

R Ge, X Feng, H Pyla, K Cameron, and W Feng. Power measurement tutorial for the green500 list. *The Green500 List: Environmentally Responsible Supercomputing*, 2007.

Eduardo C Inacio and Mario AR Dantas. Iore: A flexible and distributed i/o performance evaluation tool for hyperscale storage systems. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 01026–01031. IEEE, 2018.

Shoaib Kamil, John Shalf, and Erich Strohmaier. Power efficiency in high performance computing. In *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on*, pages 1–8. IEEE, 2008.

References II

- Athicha Muthitacharoen, Robert Morris, Thomer M Gil, and Benjie Chen. Ivy: A read/write peer-to-peer file system. *ACM SIGOPS Operating Systems Review*, 36(SI):31–44, 2002.
- Hongzhang Shan, Katie Antypas, and John Shalf. Characterizing and predicting the i/o performance of hpc applications using a parameterized synthetic benchmark. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, page 42. IEEE Press, 2008.
- Abraham Silberschatz, Greg Gagne, and Peter B Galvin. *Operating system concepts*. Wiley, 2018.
- Andrew S Tanenbaum and Maarten Van Steen. *Distributed systems: principles and paradigms*. Prentice-Hall, 2007.

References III

Parkson Wong and R Der Wijngaart. Nas parallel benchmarks i/o version 2.4. *NASA Ames Research Center, Moffet Field, CA, Tech. Rep. NAS-03-002*, 2003.