The carbon footprint of a distributed cloud storage

Lorenzo Posani[a,b], Alessio Paccoia[b], Marco Moschettini[b]

[a] *Laboratoire de Physique Statistique, École Normale Supérieure, Paris (FR)*
[b] *Research and tech development, Cubbit srl, Bologna (IT)*

Abstract

The ICT (information communication technologies) ecosystem is estimated to be responsible, as of today, for the 10% of the total worldwide energy demand - equivalent to Germany and Japan taken together. Cloud storage, mainly operated through large and densely-packed data centers, constitutes a non-negligible part of it. Moreover, every time we access a file on the cloud, a chain of routing servers need to be powered to actively transfer data from the storage facility to the users device, and back. Being a fast-inflating market and mostly an un-sensitive matter, the carbon footprint caused by "the cloud" shows no signs of slowing down. In this paper, we analyze a reversal paradigm for cloud storage (implemented by cubbit.io), in which data are stored and distributed over a network of p2p-interacting single-board ARM-based devices. We compare Cubbit' distributed cloud to the traditional centralized solution in terms of environmental footprint and power/usage efficiency. We demonstrate how a distributed architecture is beneficial for impact reduction on both sides of data transfer and data storage. Compared to centralized cloud, the distributed cloud of Cubbit achieves a $\sim$ 87% reduction of the carbon footprint for data storage and a $\sim$ 50% reduction for data transfers, providing an example of how a radical shift of paradigm can benefit both the final consumer and society as a whole.

Keywords: Carbon footprint, Cloud Storage, Distributed, Peer-to-peer

---

1. Introduction

Over the last decades, the acknowledgment of climate change by the general audience has driven the law regulations of most western countries towards an increasing awareness of consumptions and efficiency. As a result of this increasingly-tight policies, the average per-device consumption of household appliances (fridge, cooling, etc.) has consistently decreased in the last 20 years [1]. However, there is a mostly underestimated factor that sensibly contributes to the environmental impact of our daily lives: our use of the Information Communication Technology (ICT) ecosystem or, in other words, our digital life. A middle-range estimation of the total impact of the ICT ecosystem approaches 1500 TWh of annual consumption [2, 3], which roughly amounts for the 10% of the world energy consumption, more than total energy demand of Germany and Japan, taken together.

The computation of the pro-capita consumption shows that the sole fact of owning and using a smartphone constitutes, without considering the charging costs, an equivalent energy consumption of an additional operating household fridge [3]. Contrary to the electronics market, however, the footprint caused by our online life is much less tangible and, as a consequence, much less opposed. That, combined with the fast-increasing trend of online presence and internet-accessing devices pro capita, results in an increasing and mostly uncharged environmental impact that shows no signs of slowing down [4]. To gain an intuition of the impact of the digital life, one has to consider that any time a video is streamed from Youtube servers to an iPad, or a photo is accessed on Google Photos or Dropbox, the whole infrastructure that separates the final user to the corporate data center, and the data center itself, has to be powered to reliably transmit information in both directions.

The process of transmitting information can be orders of magnitude more demanding, in terms of

---

Email address: `lorenzo.posani@cubbit.io` (Lorenzo Posani)

energy consumption, than the storage itself, depending on the relative location of the exchanging nodes. In this document, we analyze, using an adaptation of the model of Baliga et al. [5], the energy consumption of *cloud storage* services, and compare it to an alternative setup where data is stored on peer-to-peer low-consumption devices located in users' houses, implemented by Cubbit, a technological startup focused on distributed cloud services.

2. Analysis of centralized cloud consumptions

The energy consumption of a cloud storage service can be divided into two main factors:

1. the cost of storing the data, i.e. powering and cooling the data center (Storage consumption)
2. the cost of sending the data from the user to the server and back (Transfer consumption)

While the first can be estimated from technical specifications of storage equipment, the second needs a more detailed analysis that takes into account the public internet infrastructure and the geographical distance between the user and the server. For both these estimations we refer to the model of Baliga et al. [5], where energy consumption is computed accounting for several factors, including the multiplicity of involved devices, redundancy, cooling, overbooking (see below).

To delineate the calculation, we start from the storage consumption, i.e. the average power, expressed in W/TB, necessary to store the payload in *hot storage*. We updated the technical specifications with respect to [5], as hard-disk storage capcity has dramatically improved in the last years. As a model for data center rack we consider the HP StoreOnce rack [6]. We take the full-operating estimation since it is the one available in the manufacturer specification sheet, and we consider full capacity (no under-usage overhead) for the 12 HDD in the rack. As an average storage per disk we consider 8 TB, estimated from the mean of HDD dimensions in blackblaze report [**?**] ($\simeq 7.2TB/Disk$). The consumption per TB is therfore estimated from the specs resumed in table 1, considering a factor 2× for cooling [**?**] and a factor 2 for redundancy. This estimation gives

$$\mathrm{P}^{\mathrm{storage}}_{\mathrm{dcenter}} = 2 \times 2 \times \frac{607\ \mathrm{kW}}{12 \times 8\ \mathrm{TB}} \simeq 25.3\ \frac{\mathrm{W}}{\mathrm{TB}} \qquad (1)$$

Similarly, we compute the transfer energy, expressed in J/GB, following the public internet model of [5]. The analysis relies on the definition of the consumption per bit, which is computed by dividing the operating power (W) for the total transfer capacity (Gb/s), resulting in a Joule/bit measure, then converted in J/GB. These units are taken from the manufacturer specs sheet, shown in table 1. These quantities are combined with a set of coefficients that reflect the redundancy of the packet transmission, the under-operating regime of the infrastructure and the cooling energy, as well as the multiplicity of some devices in a single transmission (e.g. two ethernet switches at entry points plus another one inside the data center). The average distance between core routers on the network is estimated of c.a. 800Km. For full description of coefficients and estimations we refer to [5]

$$\begin{aligned} &\mathrm{E}^{\mathrm{transfer}}_{\mathrm{dcenter}} = && (2)\\ &= 6\times\left(3\frac{P_{es}}{C_{es}} + \frac{P_{bg}}{C_{bg}} + \frac{P_g}{C_g} + 2\frac{P_{pe}}{C_{pe}} + 18\frac{P_c}{C_c} + 4\frac{P_w}{C_w}\right) \\ &\simeq 23.9\ \frac{\mathrm{kJ}}{\mathrm{GB}} \end{aligned}$$

Where the factor 6 accounts for redundancy (× 2), cooling and other overheads (× 1.5), and the fact that todays network typically operate at under 50% utilization (× 2); the addends represent, in order, the ethernet switch, the broadband gateway, the data center gateway, the provider edge router, the core network, and the relay optical fiber transmission. The detailed analysis of pre-factors can be found in [5]. Briefly, the factor 3 in the ethernet switch accounts for the two routers involved in the access to the public internet plus the router located inside the data center; the factor 18 in the core network accounts for an average of 9 hops (2 baseline + 7 for the 800km distance between core nodes) of internet packets from source to destination, times 2 for the redundancy.

3. Analysis of Distributed Cloud consumptions

The distributed architecture of the Cubbit network relies on the same public internet infrastructure delineated in the previous chapter. In the distributed paradigm, there are two key differences with the server-based cloud storage:

	Equipment	Capacity (mean)	Consumption
Storage rack powering	HP SO 3620	12 Disks x 8 TB	607 W (peak)
Cubbit Cell	Marvell ESPRESSObin	/	1 W (peak)
Cubbit Cell	WD Blue HDD	1.5 TB	1.55 W (peak)

Table 1: **Equipments** - power and capacity of storage equipments

	Equipment	Capacity	Consumption
Data Center gateway router	Juniper MX-960	660 Gb/s	5.1 kW
Ethernet Switch	Cisco 6509	160 Gb/s	3.8 kW
BNG	Juniper E320	60 Gb/s	3.3 kW
Provider Edge	Cisco 12816	160 Gb/s	4.21 kW
Core router	Cisco CRS-1	640 Gb/s	10.9 kW
WDM (800 km)	Fujitsu 7700	40 Gb/s	136 W/channel

Table 2: **Equipments** - power and capacity of routing equipments. Data from [5]

1. The low energy consumption of ARM devices (Marvell ESPRESSObin)
2. The geographical proximity between the user and his/her stored data

We consider a network of Cubbit Cells [7], each composed by an ARM-based sbc and a HDD of 1TB or 2TB (average 1.5 TB/disk). Each Cell is located in a user's house and connected by ISP internet connection. Files on the cloud are stored with a redundancy factor of 1.5 (Reed Solomon erasure coding with 24+12 redundancy shards [8, 9]). As done for the centralized cloud, we analyze the consumption of both storage (W/GB), and transfer (J/GB).

For what concerns storage, the Marvell ESPRESSObin has a single-core peak consumption of $\sim$ 1W [?]. The embedded HDD is WD Blue, that has a peak consumption of 1.4W (1 TB) and 1.7 W (2 TB). We assume that half of the network is composed by 1TB devices and half by 2TB devices, giving an average storage of 1.5TB for an average peak consumption of 1.55W. The storage energy consumption of the Cubbit network is therefore computed as

$$\mathrm{P}^{\mathrm{storage}}_{\mathrm{cubbit}} = 1.5 \times \frac{1 + 1.55\ \mathrm{W}}{1.5\ \mathrm{TB}} \simeq 2.55\ \frac{\mathrm{W}}{\mathrm{TB}} \quad . \qquad (3)$$

In Cubbit, shards of the distributed payloads are preferably distributed in Cubbit Cells that are located in geographical proximity of the user, since the distribution of the shards is controlled by the AI optimization routines of a coordinator server[8]. We therefore consider the scenario where data is stored in nodes at an average distance of 80 km from the user's access point. In this scenario, we can assume an average number of 2 packet hops in core network routers. This lowers the corresponding factor 18 in Eq. ?? to a factor 4, accounting for two core hops and the redundancy of the packets on the network (factor 2). For the same reason, the 800km-relay consumption P_w is not taken into account. With respect to Eq. ?? we also ignore the data-center-specific terms: one ethernet switch and the data center gateway. We however need to consider an additional BNG, since transfers are performed through p2p connections between endpoints located within an ISP network. The transfer energy per GB is therefore computed as

$$\mathrm{E}^{\mathrm{transfer}}_{\mathrm{cubbit}} = 6 \times \left(2\frac{P_{es}}{C_{es}} + 2\frac{P_{bg}}{C_{bg}} + 2\frac{P_{pe}}{C_{pe}} + 4\frac{P_c}{C_c} \right) \qquad (4)$$

$$\simeq 11.9\ \frac{\mathrm{kJ}}{\mathrm{GB}} \quad .$$

4. Comparison between centralized cloud and Cubbit distributed cloud

The reduction of carbon footprint of Cubbit compared to centralized solutions can be computed by comparing the storage power and the trnsfer energy for typical use-case, such as backup plans and frequent access of, for example, a web-hosted video.

By comparing the power needed to store a TB on the centralized cloud with the corresponding value for the Cubbit cloud we find

$$\Delta P^{\text{storage}} = P^{\text{storage}}_{\text{dcenter}} - P^{\text{storage}}_{\text{cubbit}} \simeq 22.75\ \frac{\text{W}}{\text{TB}} \quad , \tag{5}$$

which roughly corresponds to a 87% reduction of overall storage consumption:

$$\frac{\Delta P^{\text{storage}}}{P^{\text{storage}}_{\text{server}}} =\simeq 0.90 \quad . \tag{6}$$

Similarly, the difference in terms of transfer energy per GB is

$$\begin{aligned} \Delta E^{\text{transfer}} &= E^{\text{transfer}}_{\text{dcenter}} - E^{\text{transfer}}_{\text{cubbit}} \\ &\simeq 12.0\ \frac{\text{kJ}}{\text{GB}} = 3.33\ \frac{\text{kWh}}{\text{TB}} \quad , \end{aligned} \tag{7}$$

Which corresponds to a 50% reduction of the energy needed to transfer data from the cloud to the user, and back.

4.1. Backup

A backup service hosted on the cloud is characterized by large volumes that are not frequently accessed. In the context of the carbon footprint, the consumption of a backup plan will, therefore, be dominated by the storage term. If we consider a storage plan for a professional backup of 10 TB, with very small daily access, we find that the total saved energy in a year is

$$\begin{aligned} &\Delta E(10\ \text{TB backup}) = \\ &\quad 10\ \text{TB} \times \Delta P^{\text{storage}} \times 365 \times 24h \simeq 1992\ \text{kWh} \ . \end{aligned} \tag{8}$$

By considering a rough factor of 0.5 KgCO2 for each kWh of consumed energy[?], choosing a distributed cloud over the centralized one would correspond, for such a backup plan, to a reduced carbon emission of c.a. -1000 kgCO2/year. On a data center scale, the reduction of kgCO2 emitted per year for a PB (1000 TB) of cloud storage would therefore correspond to c.a. -100,000 kg/year.

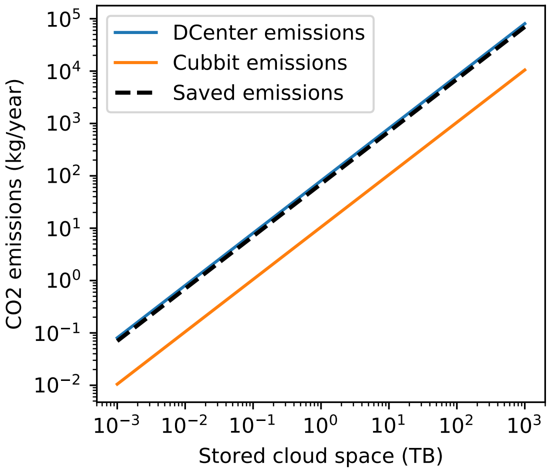


Figure 1: Comparison between centralized (Dcenter) and distributed (Cubbit) clouds in terms of annual carbon emissions per TB of stored data.

4.2. Streaming

The reduction in consumed energy, and, consequently, in carbon emission, can be significantly larger when we consider large volumes of transfer data. For example, if we consider a regional newspaper service hosting 10 TB of data and streaming, on average 10 TB of data per day (e.g. 10,000 visualizations of 100 MB each) the saved energy per year would be

$$\begin{aligned} &\Delta E(10\ \text{TB streaming}) = \\ &\quad 10\ \text{TB} \times \Delta P^{\text{storage}} \times 365 \times 24h \\ &\quad + 365 \times \Delta E^{\text{transfer}} \times 10\ \text{TB} \\ &\quad \simeq 14,100\ \text{kWh} \ , \end{aligned} \tag{9}$$

which roughly corresponds to -7,000 kgCO2 emitted per year. Note that these computations assume that streaming are broadcasted to a local audience. While this might be the case for university data, local news, or targeted marketing, it has a limited range of applicability that has to be taken into account.

4.3. Large scale

Finally, if we speculate about the overall data volume of a global consumer cloud storage service like, for example, Dropbox or Google, values rise dramatically. Such interpolations have to be taken with due caution, since estimations are based on undisclosed values. For the sake of speculation, we consider an use base of c.a. 600 millions users. The last disclosed conversion rate from free (2 GB) to

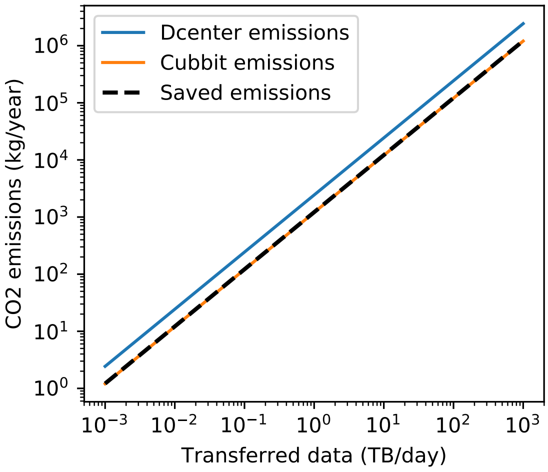


Figure 2: Comparison between centralized (Dcenter) and distributed (Cubbit) clouds in terms of annual carbon emissions per TB of daily streamed data.

pro (2 TB) is around 3%. This results in a theoretical data volume of ca. $37.2 \cdot 10^6$ TB of storage. Considering a factor 5 of overbooking, it gives an estimation of 7.4 10^6 TB of effective cloud storage. As an average daily usage, we can make a conservative estimation that each user transfers, on average, 50 MB of files from/to the cloud, which implies a daily transfer volume of c.a. 190 TB. If we plug these estimations in our model, we obtain a total saved annual energy, using a distributed architecture instead of a centralized one, of $\sim 1.5 \cdot 10^9$ kWh, equivalent to saving carbon emissions of the order of 700 millions kgCO2 per year.

Acknowledgements

The authors are grateful to Stefano Baldi (Cynny Space srl) for valuable comments and discussion.

Bibliography

[1] Energy efficiency and energy consumption in the household sector, `https://www.eea.europa.eu/data-and-maps/indicators/energy-efficiency-and-energy-consumption-5/assessment`, ????

[2] G. International, How Clean is Your Cloud? Catalysing an energy revolution, Technical Report, 2012.

[3] M. P. Mills, The cloud begins with coal, Digital Power Group (2013).

[4] Cisco vni global fixed and mobile internet traffic forecasts, `https://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html`, ????

[5] J. Baliga, R. W. Ayre, K. Hinton, R. S. Tucker, Green cloud computing: Balancing energy in processing, storage, and transport, Proceedings of the IEEE 99 (2011) 149–167.

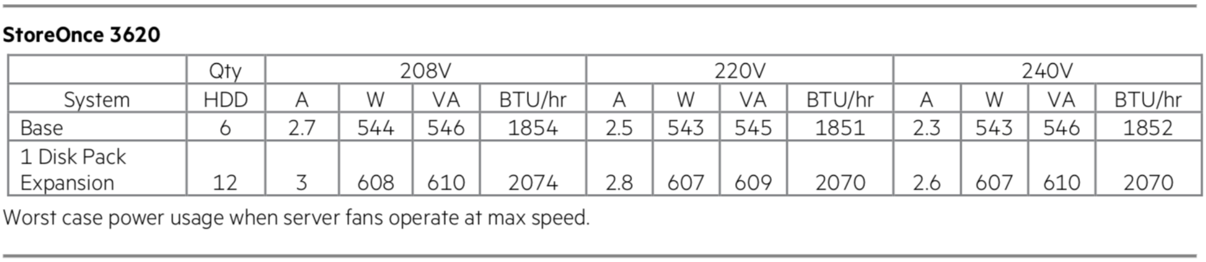
[6] Hpe storeonce data sheet, https://www.hpe.com/, ????

[7] Cubbit website, `https://www.cubbit.io/technology`, ????

[8] L. Posani, M. Moschettini, A. Paccoia, Cubbit: a distributed, crowd-sourced cloud storage service, Invited Talk at CERN Workshop on Cloud Storage Services (CS3 2018) (2018).

[9] M. Monti, S. Rasmussen, M. Moschettini, L. Posani, An alternative information plan, Technical Report, Working paper Santa Fe Institute, 2017.

Appendix: Data sheets

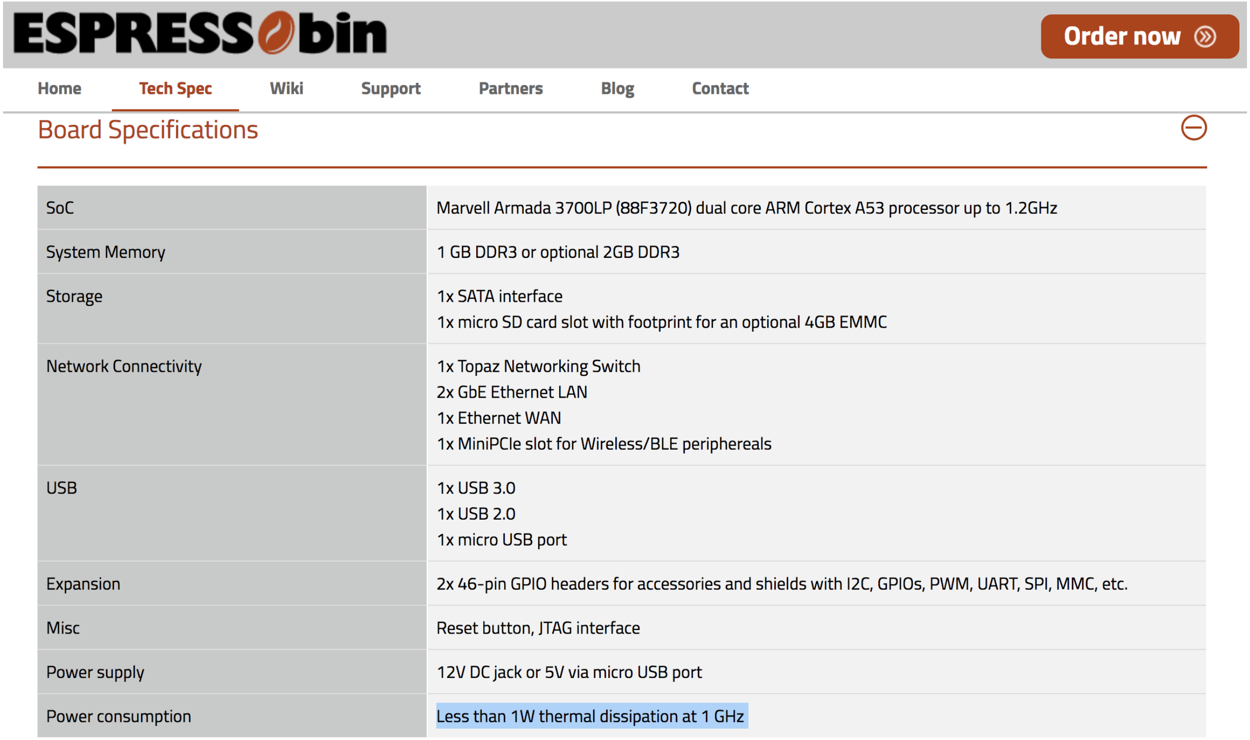


StoreOnce 3620

	Qty	208V				220V				240V			
System	HDD	A	W	VA	BTU/hr	A	W	VA	BTU/hr	A	W	VA	BTU/hr
Base	6	2.7	544	546	1854	2.5	543	545	1851	2.3	543	546	1852
1 Disk Pack Expansion	12	3	608	610	2074	2.8	607	609	2070	2.6	607	610	2070

Worst case power usage when server fans operate at max speed.

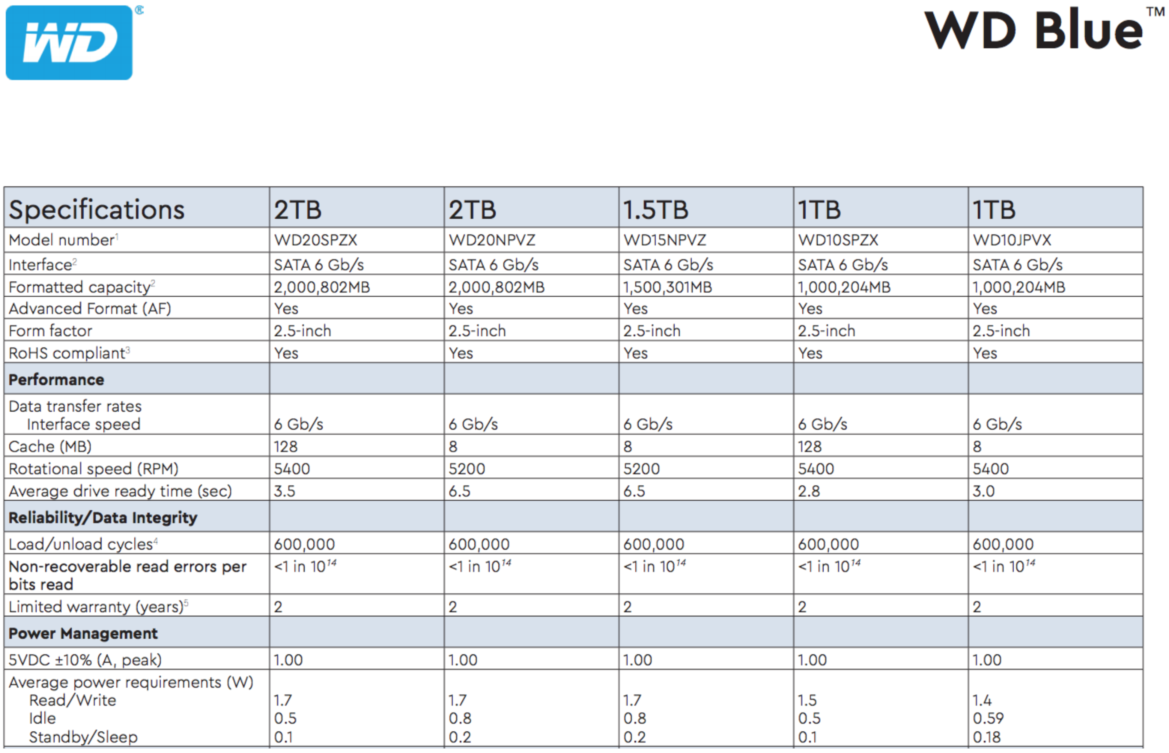
Figure 3: Appendix: HP StoreOnce 3620 specifications



Board Specifications

SoC	Marvell Armada 3700LP (88F3720) dual core ARM Cortex A53 processor up to 1.2GHz
System Memory	1 GB DDR3 or optional 2GB DDR3
Storage	1x SATA interface 1x micro SD card slot with footprint for an optional 4GB EMMC
Network Connectivity	1x Topaz Networking Switch 2x GbE Ethernet LAN 1x Ethernet WAN 1x MiniPCIe slot for Wireless/BLE periphereals
USB	1x USB 3.0 1x USB 2.0 1x micro USB port
Expansion	2x 46-pin GPIO headers for accessories and shields with I2C, GPIOs, PWM, UART, SPI, MMC, etc.
Misc	Reset button, JTAG interface
Power supply	12V DC jack or 5V via micro USB port
Power consumption	Less than 1W thermal dissipation at 1 GHz

Figure 4: Appendix: Marvell ESPRESSObin specifications



WD Blue™

Specifications	2TB	2TB	1.5TB	1TB	1TB
Model number[1]	WD20SPZX	WD20NPVZ	WD15NPVZ	WD10SPZX	WD10JPVX
Interface[2]	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s
Formatted capacity[2]	2,000,802MB	2,000,802MB	1,500,301MB	1,000,204MB	1,000,204MB
Advanced Format (AF)	Yes	Yes	Yes	Yes	Yes
Form factor	2.5-inch	2.5-inch	2.5-inch	2.5-inch	2.5-inch
RoHS compliant[3]	Yes	Yes	Yes	Yes	Yes
Performance					
Data transfer rates					
Interface speed	6 Gb/s	6 Gb/s	6 Gb/s	6 Gb/s	6 Gb/s
Cache (MB)	128	8	8	128	8
Rotational speed (RPM)	5400	5200	5200	5400	5400
Average drive ready time (sec)	3.5	6.5	6.5	2.8	3.0
Reliability/Data Integrity					
Load/unload cycles[4]	600,000	600,000	600,000	600,000	600,000
Non-recoverable read errors per bits read	<1 in 10^{14}	<1 in 10^{14}	<1 in 10^{14}	<1 in 10^{14}	<1 in 10^{14}
Limited warranty (years)[5]	2	2	2	2	2
Power Management					
5VDC ±10% (A, peak)	1.00	1.00	1.00	1.00	1.00
Average power requirements (W)					
Read/Write	1.7	1.7	1.7	1.5	1.4
Idle	0.5	0.8	0.8	0.5	0.59
Standby/Sleep	0.1	0.2	0.2	0.1	0.18

Figure 5: Appendix: Western Digital Blue HDD specifications