Accepted Manuscript

Combining edge and cloud computing for low-power, cost-effective metagenomics analysis

Daniele D'Agostino, Lucia Morganti, Elena Corni, Daniele Cesini, Ivan Merelli

PII: S0167-739X(18)30029-3
DOI: https://doi.org/10.1016/j.future.2018.07.036
Reference: FUTURE 4354

To appear in: *Future Generation Computer Systems*

Received date : 4 January 2018
Revised date : 26 April 2018
Accepted date : 17 July 2018

Please cite this article as: D. D'Agostino, L. Morganti, E. Corni, D. Cesini, I. Merelli, Combining edge and cloud computing for low-power, cost-effective metagenomics analysis, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.07.036

Combining Edge and Cloud Computing for Low-Power, Cost-Effective Metagenomics Analysis

Daniele D'Agostino[a], Lucia Morganti[b], Elena Corni[b], Daniele Cesini[b], Ivan Merelli[c]

[a] *Institute for Applied Mathematics and Information Technologies "E. Magenes", National Research Council of Italy, Genoa, Italy*
[b] *CNAF Section - Italian Institute for Nuclear Physics, Bologna, Italy*
[c] *Institute for Biomedical Technologies, National Research Council of Italy, Segrate (Milan), Italy*

Abstract

Metagenomic studies are becoming increasingly widespread, yielding important insights into microbial communities covering diverse environments from terrestrial to aquatic ecosystems. This also because genome sequencing is likely to become a routinely and ubiquitous analysis in a near future thanks to a new generation of portable devices, such as the Oxford Nanopore MinION. The main issue is however represented by the huge amount of data produced by these devices, whose management is actually challenging considering the resources required for an efficient data transfer and processing. In this paper we discuss these aspects, and in particular how it is possible to couple Edge and Cloud computing in order to manage the full analysis pipeline. In general, a proper scheduling of the computational services between the data center and smart devices equipped with low-power processors represents an effective solution.

Keywords: Metagenomics; Environmental genomics; Edge computing; Cloud computing; Internet of Things; Internet of Living Things

1. Introduction

Genome sequencing is one of the most effective analysis technique to monitor both the human body, in physiological settings and pathological conditions, as well as the bacterial communities of different environments. Developed in the 1970s with a cost of hundred million dollars, its impressive

progress [1] reduced the cost down to about $1000 dollars, and the perspective is a further reduction to about $100 for genome[1].

In particular, the MinION by Oxford Nanopore [2], a miniaturized sequencing instrument device with a weight under 100g powered by its USB port, represents one of the most promising tool belonging to the third-generation DNA sequencing technology [3]. Coupled with a laptop, MinION can be used on the field [4] to obtain genomic sequences, thus providing essential information for tracing back the organisms present in the environment [5]. These devices have been widely used for microbiology studies [2], for water monitoring[3] and in agriculture [4]. Portable sequencer can also be used to monitor bacteria in air-filters of hospitals, food industries, and pharmaceutical companies in order to give alarms in case pathogens are identified [6, 7]. More extreme usages of Oxford Nanopore devices have also been experimented [8, 9, 10].

This new trend is sometimes referred as Internet of Living Things (IoLT) [11]. The combination of the MinION sequencers with remote platform for data integration is still in its infancy, although some attempts have been reported in conference talks and blogs [12, 13]. Moreover, some prototypes of IoT platforms for monitoring clean water [14, 15], precision farming [16, 17, 18], livestock [19, 20] and more generally to improve agricultural productivity [21, 22, 23] have been presented.

The two major obstacles affecting the utility of this kind of devices are the access to suitable computational capabilities and bandwidth, since in principle it is possible to stream a couple of gigabytes of raw data per day. Even if the raw data analysis software works reliably on most of the current laptops, it is a very resource-intensive task. The adoption of a Cloud-based approach can mitigate such issue, though it comes at the expense of broadband connections.

In a previous work [24] we discussed the performance of a prototype system equipped with low-power System-on-Chip devices (SoCs). The idea is to process raw data *before*, and then send only the interesting information to a Cloud-based analysis platform able to trigger notifications and to perform

[1] https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2236383

[2] https://nanoporetech.com/publications/tags/bacteria-pathogens

[3] https://nanoporetech.com/applications#modal=water

[4] https://nanoporetech.com/applications#modal=agriculture-plant

machine learning on the time series. On the basis of the experimental results, we concluded that the pairing of a SoC and a MinION instrument (hereafter the device) allows to reduce the data streaming of 95%, thus representing a suitable solution for metagenomic analysis in remote regions.

However, this result is cost-effective for a fixed number of analyses per device per day. Nevertheless, in a real-world scenario the sampling frequency should not be fixed: whenever a critical situation is identified, the sampling is likely to be increased to monitor in more detail the evolution of the pathogens or pollution, for example to determine the effectiveness of the adopted countermeasures or the propagation in the environment. In this (realistic) case, it is important to balance the allocation of the computational services of the analysis workflow between the center and the edge of the Cloud infrastructure, in order to get timing results in a cost-effective way.

The analysis of these aspects represents the goal of the present paper, which is structured as follows. Section 2 describes the hardware components; Section 3 gives an overview of the workflow for data processing; Section 4 describes the achieved experimental results, followed by conclusions and future directions.

At the best of our knowledge, no previous works have been published about Edge computing applications in Bioinformatics. On the contrary the use of Cloud computing is a well established technology. The Cloud in fact represents a suitable solution for the storage and analysis of the present large amount of experimental data, as discussed in [25, 26]. In particular, the use of general purpose low-level solutions has been customized for the Bioinformatics research field [27, 28], as we present in this paper for the metagenomic analysis and the IoT technology. Also the price of the deployment and use of Cloud-based solutions has been discussed in several works, as [29, 30, 31]. As regards the use of SoC devices for Fog and Edge Computing applications, this approach has received increasing attention in the last years [32, 33]. This is also demonstrate by the recent (February 2018) Intel's announcement of the Xeon D-2100 line, which "brings advanced intelligence to a lower-power system-on-a-chip (SoC) for edge environments"[5].

[5]https://itpeernetwork.intel.com/xeon-taking-edge-new-heights/

2. The Hardware of the Device

A device able to operate on the field in an independent way has to be composed by MinION and a laptop/minicomputer able to manage the sampling process and its results.

The Minion device by the Oxford Nanopore is a third generation [34] approach used for sequencing DNA or RNA. Using nanopore sequencing, a single molecule of DNA or RNA can be sequenced, without the need for amplification or chemical labeling of the sample, since the molecular is able to change a ionic current passing through a nanopore. The details of this approach are described in a landmark publication [36]. Basically, they used graphene to separate two chambers containing ionic solutions and created a protein nanopore in this thin layer. The idea is that nanopore can be used as a trans-electrode, measuring a current flowing through the nanopore between two chambers. The trans electrode was used to measure variations in the current as a single molecule of DNA was passed through the nanopore. This resulted in a characteristic electrical signal that reflected the size and conformation of the DNA molecule.

A key advantage of such technology is that it makes the device portable, since it reduces the work for sample preparation. Presently this task can be accomplished in a semi-automatic way, although the company is working to make it fully automatic [35]. Moreover, MinION produces a real-time data streaming during the experiment. Indeed, this sequencing method has a capacity of 50-250 base pairs (bp) per second per pore [37]. Advantages of this method are based on the clear sequencing readout using a camera instead of noisy current methods. However, the method does require sample preparation to convert each base into an expanded binary code before sequencing. Instead of one base being identified as it translocates through the pore, 12 bases are required to find the sequence of one base [37].

These "proprietary" raw data have to be processed before their analysis with common Bioinformatics tools. In particular, there is the need to perform the *base calling operation*, which means interpreting the signals from the sequencer in order to identify the genomic sequences. The device has a declared peak throughput of 5-10 Giga base pairs (Gbp) in 48 hours, resulting in about the same amount of GB (1 base - 1 byte), even if the normal throughput is of about 0.5-2.5 Gbp in the same time interval [38]. Noteworthy, this amount of sequences can be suitable for a full metagenomic experiment, in which we want to identify bacteria that are present also in

very small amounts and with a very good precision (i.e. to identify not only their family and genus, but also their species and subspecies). On the other hand, for monitoring purpose, this accuracy is usually not necessary, since we want only to identify the presence of few specific strains. Therefore, the dataset resulting from a monitoring experiment usually has a size of about 100 Mbp, corresponding to about 30 minutes of Minion sequencing. In both cases it is clear that a broadband Internet connection is a key requirement whenever this operation is not performed locally, otherwise the portability of the device is partially impaired.

This is the reason why we investigated in [24] the use of low-power, SoC hardware platforms to equip a device including both a MinION and a dedicated minicomputer. SoCs are integrated circuits, designed for the mobile and embedded markets, composed of low power multicore processors combined with all the electronic components needed for several I/O devices. In particular, we exploited the resources of the data center of the Italian Institute for Nuclear Physics (INFN-CNAF) involved in the COSA project (COmputing On SoC Architecture[6]), an INFN initiative which aims at exploring the possibility of a greener, cost-effective and less power hungry scientific computing [39, 40]. We considered only x86-based hardware for metagenomic operations, because porting applications to these platforms is straightforward compared to other ones, i.e. ARM based, being all the dependencies already compiled and available [41, 42, 43].

In details, we investigated the use of four Intel mini-ITX boards powered by the C-2750 Avoton, the Xeon D-1540, the Pentium N3700 and the Pentium J4205 SoC CPUs. The remarkably low Thermal Design Power (TDP) of the boards, when declared, ranges from 6W of Intel Pentium N3700 to 45W of the 8-cores Intel Xeon-D processor. In an "energy-aware perspective", as shown in Figure 1 the COSA laboratory is equipped with a DC power supply, a high-precision Tektronix DMM4050 digital multimeter connected to a National Instruments data logging software, and a high-precision AC power meter, which allow to measure current and further, by integration over time, power consumptions. Table 1 provides more details about the platforms used in the present work.

All SoCs are equipped with standard 1 Gigabit Ethernet, whereas the Avoton and the XeonD are connected with both 1 and 10 Gigabit Ethernet

[6] www.cosa-project.it

Platform	Cores	Max GHz	TDP (W)	RAM (GB)	BOM (€)
Avoton C2750	8	2.4	20	16	800
XeonD 1540	8	2.6	45	16	1100
Pentium N3700	4	2.4	6	16	300
Pentium J4205	4	2.6	10	16	300

Table 1: Hardware specifications of the Intel platforms of the COSA 64bit x86 cluster hosted at INFN-CNAF in Bologna. The Bill Of Material - BOM corresponds to the money spent to acquire each platform.

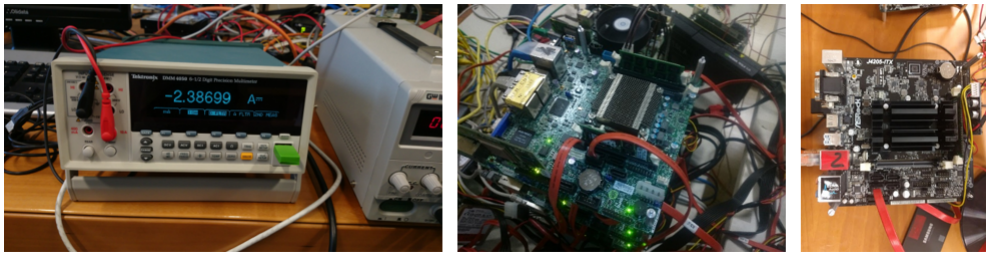

Figure 1: Pictures from the COSA laboratory, including the digital multimeter used to measure current and power consumption, and two SoCs, respectively the Avoton C2750 and the Pentium J4205.

connections. Wireless and cellular connections are also available as pluggable components.

3. The Metagenomic Analysis Workflow

A metagenomic analysis workflow relying on MinION devices consists of a variable number of operations, but the first step is in any case represented by the aforementioned base calling. In particular MinION has been designed to work with Metrichor, a Cloud-based software provided by Oxford Nanopore for performing this operation [44]. This means that the raw data have to be sent over an Internet connection for every sampling. This is the reason why many different alternative open source base callers have been developed. In particular, we identified Deepnano as one of the best solution [45]. In fact, in the last release, i.e. the R9 version, Deepnano achieves results comparable to Metrichor [46].

Most of the times, after base calling, a further operation is needed: the *bacteria identification operation*, which means classifying the type of bacteria, relying on the genomic sequences, and their relative abundance in the samples. Among the available tools we selected Kraken [47]. At the core of Kraken relies a database that contains records consisting of a k-mer annotated with the lowest common ancestor of all organisms whose genomes contain that k-mer. This database, built using a user-specified library of genomes, allows a quick lookup of the most specific node in the taxonomic tree that can be associated with a given k-mer. Each read is classified by querying the database for each k-mer in the sequence and then using the resulting set of lowest common ancestor taxa to determine an appropriate label for the read.

The results can be then managed in a domain-specific way. As stated in the Introduction, we considered two kinds of applications:

- the triggering of alarms in case specific and dangerous bacteria are found in the air-filter (e.g. of hospitals, food industries, and pharmaceutical companies) or if pollution signals reach a given threshold (e.g of rivers, lakes, aqueducts);

- the identification and analysis of set points, through data integration and machine learning, of the bacterial communities in soil (e.g. for cultivations, greenhouses and animal farms), providing feedback in case of deviations.

These two analysis services can be executed in the Cloud using one of the IoT and data analytic platforms available. The results of the bacteria identification operation in fact have a size of a few MB, therefore their upload does not represent an issue.

Presently, there is an increasing number of commercial platforms (e.g. AWS IoT, Microsoft Azure IoT, Google Cloud Platform, IBM Watson, Intel IoT) and research projects (e.g., OpenIoT, Waziup, Kaa) that can be effectively exploited. They differ on many aspects as easiness of use, supported languages, security management, available integrated solutions, performance and cost efficiency [48]. In general, it is not possible to identify a single solution that perfectly addresses the needs of applications and developers [49]. Therefore, we selected the AWS IoT platform since it is straightforward to use and it provides a rich, integrated environment, e.g. the machine learning platform services, that suits our requirements.

AWS IoT is a cloud managed platform providing a publish/subscribe brokering service. It offers out of the box a number of features as security and the seamless integration with the AWS service ecosystems, like Lambda functions, DynamoDB, S3 and many more. The platform basically enables the bi-directional communication between Internet-connected *things* (e.g. sensors and applications) through logical channels. The communication is based on JSON messages addressing topics like *minion/location_xy*. A message broker sends the message to all clients that have registered to receive messages for a topic. The act of sending the message is referred to as *publishing*. The act of registering to receive messages for a topic filter is referred to as *subscribing*.

The simplest architecture that can be implemented with AWS IoT is composed only by *things* and the device gateway. In fact, an external Web service can register to the device gateway, subscribe to all the topics defined in the applicative scenario and manage them in an independent way, i.e. for storing them on a proprietary repository or for taking actions as sending alerts. Otherwise the Rules Engine has to be activated. The difference of having a Web service subscribing a topic or the use of an action relies on the scalability AWS can provide. Our prototype exploits the Rules Engine to filter the messages, while the actual alarm triggering is managed through a dedicated Web services. Also the storing of results was performed with local resources.

This architecture is shown in Figure 2. The most demanding analysis services in terms of compute capabilities sit on the Edge of the Cloud, while the collection of the results of all the devices and the following application-specific processing is performed on the Cloud. The advantage is represented by having a scalable system in terms of the number of devices, but not in the number of sampling performed by each single device.

In a real-world scenario, the sampling frequency in fact is likely to be dynamically determined, e.g. whenever a critical situation arises. With MinION we can consider to perform up to 20 sampling per day per device. In this case the use of only one local SoC is not sufficient. Therefore three alternative strategies are possible to perform the base calling and bacteria identification services, namely:

- to equip the device with more compute capabilities;
- to move them on the Cloud infrastructure;
- to rely on a Fog-based solution.

These scenarios arise from the consideration that devices are likely to have heterogeneous conditions in terms of connectivity. For example, Figure 2 shows the monitoring network for surface water in Lombardy, composed by more than 400 stations[7]. Lombardy is the richest Italian department in lakes - about 50, representing 40% of the national surface total. Moreover the overall length of rivers and irrigation channels reaches about 200,000 km, which support the agricultural activities. Due to the vast urbanization of the territory, the industrialization and the diffusion of agro-zootechnical activities, the water resources need constant monitoring and protection measures[8]. The present distribution is the result of many considerations but also constraints, among them the access to suitable connectivity and energy supply facilities. Therefore the adoption of an architecture and technological solutions like those discussed here can overcome such constraints and permit to evaluate a different positioning and an increase in the number of the monitoring station.

In the following Section we discuss some quantitative tests to evaluate the different strategies for designing a possibly ubiquitous system for environmental metagenomic analysis. The focus is to assess how many resources (mainly the computational capabilities and the network bandwidth) are required to support the different scenarios, their feasibility and cost-efficiency.

4. Experimental Results

As discussed before, the sensitivity of an experiment heavily depends on the depth of sequencing [50]. While MinION can produce in theory for a full discovery metagenomic experiment up to 2 Gbp in 20 hours, 100 - 150 Mbp in about 1 hour are sufficient for monitoring purpose [51]. Moreover MinION streams the sequences as a set of files that can be processed as soon as they are available.

The monitoring analysis represents the test case we consider hereafter. Therefore we can figure to increase the sampling frequency up to 20 sampling per day. This behavior can be implemented in a straightforward way by deploying a Web service on the devices that periodically receives this value by the Cloud-based monitoring system.

As dataset for our tests we considered a set of files, with a global size of 100 MB, derived from the sequencing of a metagenomic experiment [52].

[7] http://www.arpalombardia.it/sites/arpalombardia2013/RSA/Pagine/tematismo.aspx?p1=2145
[8] http://www.lambrovivo.eu/?page_id=17

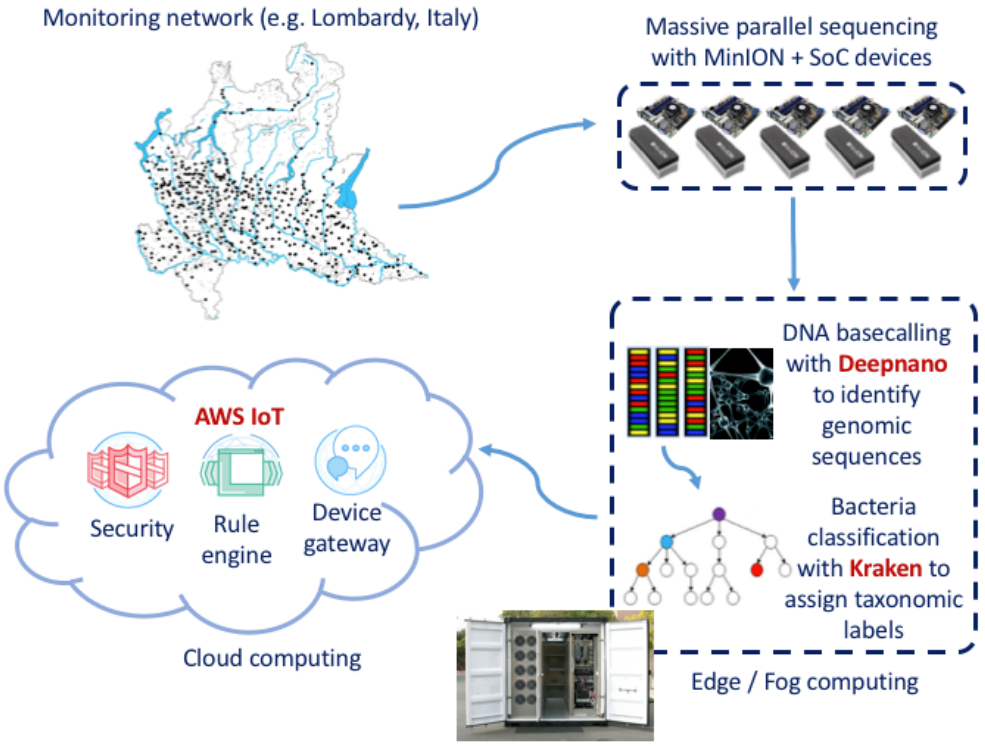

Figure 2: The architecture of the system for metagenomic analysis based on MinION + SoC devices and possible applicative scenarios.

They are processed by Deepnano, that produces the same amount of files, with a global size of 4.6 MB in the FASTA format, to be processed by Kraken. The final result is represented by 790 tuples, with a global size of 3.2 MB.

Table 2 shows the amount of bases per second that the considered SoCs can process for Deepnano and Kraken, in case a single core or all the available cores are used. In this respect, it is to note that, given multithreading is currently not supported, Deepnano has been executed in a data-parallel way by splitting the dataset and running independent instances of the application, one for each core available in the SoC. Moreover, the analysis of which bacteria are present and in which amount has been performed with the MiniKraken version, supplied specifically for low-memory computing environments. Kraken in fact requires at least 75 GB to hold its database in RAM. MiniKraken instead uses a reduced version of the database because it considers only the relevant k-mers in a sequence to get the correct classification, a procedure which does not invalidate the analysis.

	DEEPNANO [bps]		KRAKEN [bps]	
	1 core	All cores	1 core	All cores
XeonD	367.9	1801.9	$2.4 * 10^6$	$9.9 * 10^6$
Avoton	152.85	1225.3	$1.1 * 10^6$	$4.0 * 10^6$
N3700	216.5	609.9	$0.9 * 10^6$	$1.9 * 10^6$
J4205	111.1	237.5	$1.4 * 10^6$	$2.9 * 10^6$

Table 2: Bases processed per second by Deepnano and Kraken for the different low-power architectures considered in this work, using one or all the available cores.

If interested in one sampling per day for monitoring purposes, i.e. in processing 100 - 150 Mbp, the results in Table 2 show that only the XeonD and Avoton are able to process it, whereas for the other tested platforms there is the need to consider 2-5 boards in order to get results in time. To provide a comparison figure, the Xeon E5-2683 v3 CPU (14 Cores, 3.0 GHz Max, TDP 120W, BOM 3000 €), used in a High Performance Computing (HPC) cluster representing the reference for the COSA project, is able to process about 86 Mbp per day with a single core and up to 1.2 Gigabases using all its 14 physical cores. Considering the energy consumption, a single

analysis requires 1.9 MJoule with the XeonD, 2 MJ with the Avoton and 18.8 MJ with the Xeon E5.

When more than a sampling per day is performed, the use of only the local, single SoC CPU is not sufficient. Therefore we can either equip the device with more SoC CPUs or equip the device with HPC CPUs.

Using more SoC CPUs

We have to equip the device with 14 XeonD or 20 Avoton CPUs in order to be able to support the highest rate of 20 sampling per day. This solution is however the worst because:

1. the cost: considering the price of each SoC platform, each device will cost more than 14,000 €;
2. the usage of resources: most of the time, only one SoC CPU will be active because a single analysis will be performed;
3. the power consumption: about 27 MJ per day are necessary to operate all the SoCs;
4. the size of the resulting device: a box containing 14 mini-ITX motherboards (15 cm size each) resembles about the size of four shoeboxes.

Using HPC CPUs

In this scenario, we have to equip the device with 2 Xeon E5-2683, able to process the 2 Gbases resulting from the highest sampling frequency. This solution is suitable only for not battery-powered devices, because a single day of analysis requires about 380 MJ. However this solution has a lower cost with respect to the previous one, i.e. 6,000 €, even if also in this case the CPUs are underutilized for most of the time.

In general, both these solutions are not sufficiently cost-effective, thus we can conclude that when more than one analysis per day is required it is better to move the analysis services closer to the center of the Cloud.

Using Cloud computing services

The key issue of moving all the computational services on the Cloud when the sampling frequency is more than one is represented by the size of the data to be sent over a wired or wireless connection. The main advantage of the SoC-equipped device is in fact represented by the reduction of the data to be transferred from 100 down to 5 MB. We experimented the use of several

Tool	Avoton Exec. time (s)	XeonD Exec. time (s)	Resulting Size (MB)
gzip	6.9	4.4	92.9
bzip2	33.7	17.6	92.6
xz	62.2	37.2	91.6
pigz	0.9	0.7	93.1
pbzip2	5.9	2.5	92.6
pxz	15.1	9.7	91.7
7z	9.3	4.9	92.4

Table 3: Execution times, in seconds, and resulting size of compressed data, considering a set of files of 100 MB. While gzip, bzip and xz are sequential programs, the others can be run as multithreaded applications.

compression algorithms, both sequential and parallel. However they are not able to reduce considerably the size, as shown in Table 3.

This means that the device should be able to use a wired or wireless connection providing at least 215 Kbps in uplink of actual user data in order to be able to send the best result, 91.7 MB, in about one hour, i.e. when the next sampling result is available. In fact, the possibility to send 91.7 MB implies a bandwidth of at least 26 Kilo Byte per second, i.e. 208 Kilo bit per second. While this number is not an issue with wired and broadband wireless connections, the availability of only 2.5/3G technologies (i.e. GPRS) represents a major limit. GPRS connections can offer in fact as little as 20 Kbps. The theoretical limit of 171 Kbps is never realized on modern networks - 50 Kbps download and 25Kbps upload is the typical data transfer rate available. Also the 3G-UMTS networks do not offer a sufficient bandwidth, i.e. 384 Kbps download and 64 Kbps upload. In the latter case, however, it is possible to implement multiple data streams using multiple SIM cards and antennas, as it is a common solution for video streaming.

The power consumption of sending the data is rather negligible: most of the wired/wireless cards requires in fact 1 J. Hence, about 3.6 kJ are necessary to send a single analysis, and 72 kJ for a single day. In case of mains powered devices is also possible to use the powerline networking, thus

reducing the consumption to a negligible value.

Therefore this solution is the best one for devices under the coverage of high-speed networks.

Using Fog computing services

This last solution represents a middle ground among the previous ones, and it consists of providing a computational node on the LAN represented by a set of N devices following the Fog computing paradigm [53, 54]. For example it is possible to exploit portable datacenters[9] when critical situations occur.

An important aspect for the cost-effectiveness is a careful evaluation of the compute capability of the node, by considering the probability that M $\leq$ N devices are running analysis at a time at the full speed. The advantages of such solution can be summarized as follows:

1. the cost and the usage: only 2M Xeon CPUs are necessary with respect to the 2N required by local solutions, with a cost of 6,000*M instead of 6,000 * N;
2. the possibility to deploy the node only when and where it is necessary.

Furthermore, the transmission time does not represent an issue any more because an ad-hoc network can be provided [55, 56, 57, 58] for the heavy communication among the devices and the node, keeping low the size of the final results to be sent from the node over the Internet connection. Other solutions relying on long-range wifi infrastructures [59, 60, 61] are possible.

This last strategy therefore represents the best one for battery-powered devices under the coverage of low-bandwidth networks.

5. Conclusion and Future Development

Metagenomic studies are becoming increasingly widespread, yielding important insights into microbial communities covering diverse environments from terrestrial to aquatic ecosystems. With the advent of high-throughput sequencing platforms, the use of large scale shotgun sequencing approaches is now commonplace.

[9] https://www.ibm.com/us-en/marketplace/prefabricated-modular-data-center

In a previous work we discussed an architecture and the performance of a prototype based on low-power Systems-On-Chip for metagenomic analysis able to support a fixed number of routinely analysis per day. In this paper we presented an evolution of such architecture, which supports the possibility to dynamically increase or decrease the sampling rate when critical situations occur.

We analyzed four different strategies and we concluded that, while the previous architecture is an effective solution when a single analysis per device is performed every day, the best solution when the frequency increases - considering both cost and performance - is to "move" computational services from the Edge to the Fog or Cloud infrastructures, depending on the available Internet connection.

We plan to extend this work in two ways. First, we will analyse the use of ad-hoc/long-range wifi networks for deploying wider distributed system relying on the Fog computing paradigm.

Second, we will design more complex data integration systems, to aggregate results from different devices, and machine-learning approaches, in order to identify some set points in the microbial composition providing optimal results in the farm/industrial production, for example in milk productions or smart agriculture. Our prototype in fact represents a solution addressing the problem of managing networks of IoT devices producing large datasets before sending data on a Cloud environment using possibly low-bandwidth networks. Once the data have been transferred, it is possible to develop a specific solution, calibrated ad hoc for each considered scenario, with the aim to provide feedback information to maintain the identified set points on the long period.

Acknowledgement

The Minion Oxford Nanopore technology was available to the project thanks to Pietro Lio, University of Cambridge, who was involved in the first Oxford Nanopore MinION Access Program. The accessibility to the AWT platform was granted by a AWS Cloud Credits for Research to Ivan Merelli. This work has been supported by the Italian flagship initiative Interomics for CNR researchers and by the INFN COSA Project for CNAF-INFN researchers.

References

[1] Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends in genetics, 30(9), 418-426.

[2] Jain, M., Olsen, H. E., Paten, B., Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology, 17(1), 239.

[3] Feng, Y., Zhang, Y., Ying, C., Wang, D., Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. Genomics, proteomics & bioinformatics, 13(1), 4-16.

[4] Walter, M.C., Zwirglmaier, K., Vette, P., Holowachuk, S.A., Stoecker, K., Genzel, G.H., Antwerpen, M.H. (2017), MinION as part of a biomedical rapidly deployable laboratory, Journal of Biotechnology, 250, 16.

[5] Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C., Franklin, R.B. (2017), MinION nanopore sequencing of environmental metagenomes: a synthetic approach, Gigascience, 6(3), 1.

[6] Alsan, M., and Klompas, M. (2010), Acinetobacter Baumannii: An Emerging and Important Pathogen. Journal of clinical outcomes management, JCOM 17.8, 363.

[7] Peleg, A.Y., and Hooper, D.C. (2010), Hospital-Acquired Infections Due to Gram-Negative Bacteria. The New England journal of medicine 362.19, 1804.

[8] Edwards, A., Debbonaire, A.R., Sattler, B., Mur, L.AJ, Hodson, A.J. (2016), Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N, Biorxiv.

[9] Castro-Wallace, S.L., Chiu, C.Y., John, K., Stahl, S.E., Rubins, K.H., McIntyre, A.B.R., Dworkin, J.P., Lupisella, M.L., Smith, D.J., Botkin, D.J., Stephenson, T.A., Juul, S., Turner, D.J., Izquierdo, F., Federman, S., Stryke, D., Somasekar, S., Alexander, N., Yu, G., Mason, C., Burton, A.S. (2016), Nanopore DNA Sequencing and Genome Assembly on the International Space Station, Biorxiv.

[10] McIntyre, A.B.R., Rizzardi, L., Yu,A.M., Alexander, N., Rosen, G.L., Botkin, D.J., Stahl, S.E., John, K., Castro-Wallace, S.L., McGrath, K., Burton, A.S., Feinberg, A.P., Mason, C.E. (2016), Nanopore Sequencing in Microgravity, npj Microgravity 2, Article number: 16035.

[11] Pauwels, E. (2017) The Internet of Living Things. Scientific American. https://blogs.scientificamerican.com/observations/the-internet-of-living-things/

[12] Waltz, E. (2016) Portable DNA Sequencer MinION Helps Build the Internet of Living Things. https://spectrum.ieee.org/the-human-os/biomedical/devices/portable-dna-sequencer-minion-help-build-the-internet-of-living-things

[13] Medeiros, J. (2016) DNA analysis will build an internet of living things. http://www.wired.co.uk/article/dna-analysis-internet-living-things

[14] BoonFei, T. et al. (2015), Next-Generation Sequencing (NGS) for Assessment of Microbial Water Quality: Current Progress, Challenges, and Future Opportunities, Frontiers in Microbiology 6, 1027.

[15] Rahim, A. (2017). IoT and Data Analytics for Developing Countries from Research to Business Transformation. In International Conference on the Economics of Grids, Clouds, Systems, and Services (pp. 281-284). Springer, Cham.

[16] Bouloukakis, M., Stratakis, C., Stephanidis, C. (2018) AmI Garden: Building an IoT Infrastructure for Precision Agriculture. ERCIM News 113, pp. 18-19.

[17] Lovas, R., Koplanyi, K., Elo, G. (2018) Agrodat: A Knowledge Centre and Decision Support System for Precision Farming Based on IoT and Big Data Technologies. ERCIM News 113, pp. 22-23.

[18] Rolf, D. (2005), The metagenomics of soil. Nature Reviews Microbiology, 3 pp. 470478.

[19] Ganda, E.K., et al. (2016), Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin. Scientific reports, 6.

[20] Shanks, O. C., Kelty, C. A., Archibeque, S., Jenkins, M., Newton, R. J., McLellan, S. L., ... Sogin, M. L. (2011). Community structures of fecal bacteria in cattle from different animal feeding operations. Applied and Environmental Microbiology, 77(9), 2992-3001.

[21] Boschetti, M., Schoitsch, E. (2018) Smart Farming. ERCIM News 113, pp. 16-17.

[22] Rolf, D. (2005), The metagenomics of soil. Nature Reviews Microbiology, 3.6.

[23] Castaneda, L.E., and Barbosa, O. (2017), Metagenomic Analysis Exploring Taxonomic and Functional Diversity of Soil Microbial Communities in Chilean Vineyards and Surrounding Native Forests, Ed. Keith Crandall. PeerJ 5, e3098.

[24] Merelli, I., Morganti, L., Corni, E., Cesini, D., Roverelli, L., Zereik, G., D'Agostino, D. (2018), Low-Power Portable Devices for Metagenomics Analysis: Fog Computing Makes Bioinformatics Ready for the Internet of Things. Future Generation Computer Systems, vol. 88, pp. 467-478. .

[25] Dai, L., Gao, X., Guo, Y., Xiao, J., Zhang, Z. (2012), Bioinformatics clouds for big data manipulation, Biol Direct, 7, 43.

[26] Merelli, I., Perez-Sanchez, H., Gesing, S., DAgostino, D. (2014), Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives, BioMed research international.

[27] Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D., Nelson, K.E. (2012), Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community, BMC Bioinformatics, 13, 42.

[28] Afgan, E., Chapman, B., Jadan, M., Franke, V., Taylor, J. (2012), Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy. Curr Protoc Bioinformatics, 11.9.

[29] D'Agostino, D., Clematis, A., Quarati, A., Cesini, D., Chiappori, F., Milanesi, L., Merelli, I. (2013), Cloud infrastructures for in silico drug discovery: economic and practical aspects, BioMed research international 2013.

[30] Guerrero, G. D., Wallace, R. M., VazquezPoletti, J. L., Cecilia, J. M., Garcia, J. M., Mozos, D., PerezSanchez, H. (2014). A performance/cost model for a CUDA drug discovery application on physical and public cloud infrastructures. Concurrency and Computation: Practice and Experience, 26(10), 1787-1798.

[31] Guerrero, G. D., Imbernon, B., Perez-Sanchez, H., Sanz, F., Garcia, J. M., Cecilia, J. M. (2014). A performance/cost evaluation for a GPU-based drug discovery application on volunteer computing. BioMed research international, 2014.

[32] Arnab Kumar, B., Nandy, S.K., Narayan, R. (2017), Multiprocessor system-on-chip for processing data in cloud computing, Data Security in Cloud Computing, 65.

[33] Conti, F. et al. (2017), An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics, IEEE Transactions on Circuits and Systems I, 64(9), 2481.

[34] Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., Barron, A. E. (2011). Landscape of next-generation sequencing technologies. Analytical chemistry, 83(12), 4327-4341.

[35] Ko, S. Y., Sassoubre, L., Zola, J. (2017). Applications and Challenges of Real-time Mobile DNA Analysis. arXiv preprint arXiv:1711.07370.

[36] Garaj, S., Hubbard, W., Reina, A., Kong, J., Branton, D., Golovchenko, J. A. (2010). Graphene as a subnanometre trans-electrode membrane. Nature, 467(7312), 190.

[37] McNally, B., Singer, A., Yu, Z., Sun, Y., Weng, Z., Meller, A. (2010). Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. Nano letters, 10(6), 2237-2244.

[38] Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J. (2015), Assessing the performance of the Oxford Nanopore Technologies MinION, Biomolecular Detection and Quantification, 3, 1.

[39] Cesini, D., Corni, E., Falabella, A., Ferraro, A., Morganti, L., Calore, E., Schifano, S.F., Michelotto, M., Alfieri, R., De Pietri, R., Boccali,

T., Biagioni, A., Lo Cicero, F., Lonardo, A., Martinelli, M., Paolucci, P.S., Pastorelli, E., Vicini, P. (2017), Power-Efficient Computing: Experiences from the COSA project, Scientific programming, Article ID 7206595.

[40] D'Agostino, D., Cesini, D., Corni, E., Ferraro, A., Morganti, L., Quarati, A., Merelli, I. (2017), Performance and Economic Evaluations in Adopting Low Power Architectures: A Real Case Analysis. In International Conference on the Economics of Grids, Clouds, Systems, and Services (pp. 177-189). Springer, Cham.

[41] Morganti, L., Cesini, D., Ferraro, A. (2016, February). Evaluating Systems on Chip through HPC bioinformatic and astrophysic applications. In Parallel, Distributed, and Network-Based Processing (PDP), 2016 24th Euromicro International Conference on (pp. 541-544). IEEE.

[42] Corni, E., Morganti, L., Morigi, M. P., Brancaccio, R., Bettuzzi, M., Levi, G., ... Ferraro, A. (2016, February). X-Ray computed tomography applied to objects of cultural heritage: porting and testing the filtered back-projection reconstruction algorithm on low power systems-on-chip. In Parallel, Distributed, and Network-Based Processing (PDP), 2016 24th Euromicro International Conference on (pp. 369-372). IEEE.

[43] Morganti, L., Corni, E., Ferraro, A., Cesini, D., D'Agostino, D., Merelli, I. (2017, March). Implementing a space-aware stochastic simulator on low-power architectures: a systems biology case study. In Parallel, Distributed and Network-based Processing (PDP), 2017 25th Euromicro International Conference on (pp. 303-308). IEEE.

[44] Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., Akeson, M. (2015), Improved data analysis for the MinION nanopore sequencer, Nature methods, 12(4), 351.

[45] Boza, V., Brejova, B., Vinar, T. (2017), DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads, PloS one, 12(6), e0178751.

[46] de Lannoy, C., de Ridder, D., Risse J. (2017), A sequencer coming of age: De novo genome assembly using MinION reads. F1000Research 2017, 6:1083.

[47] Wood D.E., Salzberg, S.L. (2014), Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biology, 201415:R46.

[48] Botta, A., De Donato, W., Persico, V., Pescap, A. (2016). Integration of cloud computing and internet of things: a survey. Future Generation Computer Systems, 56, 684-700.

[49] Ray, P. P. (2016). A survey of IoT cloud platforms. Future Computing and Informatics Journal, 1(1-2), 35-46.

[50] Jiajia, N., Qingyun, Y., Yuhe, Y. (2013), How much metagenomic sequencing is enough to achieve a given goal?. Scientific reports, 3.

[51] Lindgreen, S., Adair, K.L., Gardner, P.P. (2016), An evaluation of the accuracy and speed of metagenome analysis tools, Scientific Reports 6, Article number: 19233.

[52] Brown, B. L., Watson, M., Minot, S.S., Rivera, M. C., Franklin, R.B. (2017), MinION nanopore sequencing of environmental metagenomes: a synthetic approach. GigaScience, 6(3), 1.

[53] Dastjerdi, A.V., Rajkumar Buyya, R. (2016), Fog Computing: Helping the Internet of Things Realize Its Potential. Computer 49(8), 112.

[54] Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are. CISCO Whitepaper, 2015

[55] Bruzgiene, R., Narbutaite, L., Adomkus, T. (2017). MANET Network in Internet of Things System. Chapter 5 in Ad Hoc Networks. InTech.

[56] Reina, D.G., Toral, S.L., Barrero, F., Bessis, N., Asimakopoulou, E. (2013), The Role of Ad Hoc Networks in the Internet of Things: A Case Scenario for Smart Environments, Chapter 4 in Internet of Things and Inter-cooperative Computational Technologies for Collective Intelligence, 89.

[57] Mukherjee, S., Biswas, G.P. (2017), Networking for IoT and applications using existing communication technology, Egyptian Informatics Journal, online.

[58] Qin, H., Chen, W., Cao, B., Zeng, M., Peng, Y. (2018), A cross-interface design for energy-efficient and delay-bounded multi-hop communications in IoT, Ad Hoc Networks, 70, 103.

[59] Chebrolu, K., Raman, B., Sayandeep, S. (2006), Long-distance 802.11b links: performance measurements and experience, Proceedings of the 12th annual international conference on Mobile computing and networking (MobiCom '06), 74.

[60] Cerda-Alabern, L., Neumann, A., Escrich, P. (2013), Experimental evaluation of a wireless community mesh network, Proceedings of the 16th ACM international conference on Modeling, analysis & simulation of wireless and mobile systems (MSWiM '13), 23.

[61] Vega, D., Baig, R., Cerda-Alabern, L., Medina, E., Meseguer, R., Navarro, L. (2015), A technological overview of the guifi.net community network, Computer Networks, 93(2), 260.

Daniele D'Agostino, Ph.D., is a researcher at the Institute of Applied Mathematics and Information Technologies of National Research Council. His research activities concern the design of science gateways in different research fields, the resource allocation in Grid/Cloud environments and the development of parallel software. He co-organized the 22th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, several special issues on ISI journals and co-authored more than 80 scientific papers, published in journals, book chapters and conference proceedings.

WEB: http://www.imati.cnr.it/index.php/it/daniele-d-agostino

Lucia Morganti, PhD in Astrophysics from the Ludwig Maximilians University of Munich, works at CNAF, the INFN National Center dedicated to research and development on IT technologies, located in Bologna, Italy. She is a member of both the Storage and the User support unit of the data center, and she is involved in the H2020 ExaNeSt project for exascale-level supercomputers, and in the INFN-COSA (Computing On SoC Architecture) project for power-efficient scientific computing.

WEB: https://www.cnaf.infn.it/~lmorganti/web/home.html

Elena Corni, master degree in Medical Phisycs at University of Bologna, works at INFN-CNAF in Bologna, Italy. She is a member of the User Support unit of the data center, and she collaborates in the INFN-COSA (Computing On SoC Architecture) project.

WEB: https://www.linkedin.com/in/elena-corni-8b35a6b6/

Daniele Cesini is working as a Researcher in Technology at the Italian Institute for Nuclear Physics (INFN). He is currently a member of the Data Handling group at INFN-CNAF. He is the coordinator of the User Support Team of the INFN Tier1. Since 2004, he acquired experience working within national and international initiatives dealing with distributed and parallel computing. He focused his research in the field of efficient tasks-scheduling in distributed environments for mixed High Performance/High Throughput Computing workflows. He is expert in the application porting to different computing platforms: distributed architectures, low power processors and HPC hybrid systems.

WEB: https://www.cnaf.infn.it/author/cesini/

Ivan Merelli is staff scientist at the Institute for Biomedical Technologies (ITB) of the Italian National Research Council (CNR) in the Bioinformatic Unit. He earned a PhD in computer science from the University of Milano-Bicocca in 2009 and was visiting scientist at Harvard University in 2014 and at Cambridge University in 2015. He got the Italian associate professor habilitation in 2017. He coauthored more than forty papers published in international peer-reviewed journals and more than forty contributions in International Conference Proceedings. His research activities concern statistical data analysis, software development and data management in the field of *Genomics* and *Proteomics*, in particular for the management of *Biological Databases* and *High Performance Computing* facilities. He participated to the Italian Projects Grid.it, LITBIO and ITALBIONET and to the European Projects BioinfoGRID and EGEE for the development of high performance solutions in the field of Systems and Computational Biology. He is associated editor of of "BMC Genomics", BioMed Central Ltd and "Frontiers in Bioinformatics and Computational Biology", Frontiers Media S.A. He was Guest Editor of the special issue on "Latest advances in distributed, parallel, and graphic processing unit accelerated approaches to computational biology" in "Concurrency and Computation: Practice and Experience" and of the special issue "High-Performance Computing and Big Data in Omics-Based Medicine" in "Biomed Research International".

WEB: http://www.cnr.it/people/ivan.merelli

Portable sequencing machines can be used for monitoring the microbioma in different environments

Low-power devices can be used to analyze sequencing data in real-time on the field

Cloud IoT platforms can be used to trigger alarms (rules) or to identify set points (data analytics)

A proper coupling of Edge and Cloud computing is necessary in real-world scenario, where the sampling frequency is dynamically determined