Architecture Considerations for Big Data Management

Khalim Amjad Meerja
Western University, London Ontario,
Canada
Email: kmeerja2@uwo.ca

Khaled Almustafa
Prince Sultan University,
Riyadh, K.S.A.
Email: kalmustafa@psu.edu.sa

***Abstract*—A network architecture is concerned with holistic view of interconnection of different nodes with each other. This refers to both physical and logical ways of interconnection of all nodes in the network. The way in which they are connected influences the strategies adopted for Big Data Management. In this present day of Internet of Things (IoT), each kind of device is required and made possible for communicating with other completely different kinds of devices. The heterogeneous nature of devices in the network needs a completely new architecture to efficiently handle Big Data which is generated continually, either for providing services to end users or for study and analysis in a research process. It is thus very essential to visit various kinds of devices that are available on the Internet, their characteristics and requirements, how they communicate and process data, and eventually how the human society embraces the Big Data generation for their daily consumption. This paper is dedicated to bringing all theses aspects together in one place, bringing different technologies into one single network architecture.**

***Keywords*—Internet of Things (IoT), Big Data, Cloud network, RFID, Sensor Networks, 5G.**

I. Introduction

A notable feature of our present day network communications is new ways in which different kinds of devices talk to each other over the Internet. Devices vary in shapes and sizes. Some are from very tiny to small, operating autonomously while others are comparatively large and more complex. They are highly heterogeneous in many aspects, having different functionalities, computational capabilities, and storage. With regards to mobility, some are highly mobile, while others are either portable or fixed in their locations. They work in master slave relation or as peers. They transmit multimedia such as voice, video, images, and text. All these various functionalities need different network conditions and communication practices to operate better and serve.

For instance, tiny devices have constrained energy resources. Their operations are based on energy conservation mechanisms and needs completely different practices such as network protocols, architectures, energy harvesting mechanisms, robustness to failures in adverse conditions, system security measures, to name a few. Data has to be periodically collected and forwarded to more stable and powerful system for manipulation and analysis. Larger devices such as smart computational devices have higher expectations in terms of services offered. So, their underlying network conditions have to be more powerful and robust. It is therefore important to look into the available types of devices and their network requirements. From studying all these network requirements, it is possible to come up with an unified architecture as a solution to efficiently handle Big Data evolving from trillions of such devices connected to the Internet.

A. Wireless Sensor Networks

Wireless sensor networks are predominantly used for environmental, industrial, agricultural, health, and habitat monitoring purposes [1]. These are mostly ad hoc networks that were originally developed for communication in military operations. Later they are found in many useful applications in domestic environment [2]. Sensor devices are generally placed in harsh environmental conditions. Therefore node failures are common in sensor networks. Network routing and media access control protocols should deal with node failures and still maintain communication links with their destination sink nodes. While doing so, they have to conserve their energy as it is difficult to replenish them with power sources. In some cases, it is possible that these sensor devices derive energy from solar cells or from some sort of mechanical stress induced by onboard piezoelectric material.

Network architecture should consider all these issues. More importantly, fault tolerant mechanisms are essential to keep up with node failures. Some node failures are intermittent and short lived while some are permanent. The path between the network nodes will constantly be changing due to these reasons and also because of another important reason, which is the unreliable nature of wireless channels. The sink nodes have to maintain a simultaneous connection with a group of sensor nodes in a particular region that are performing monitoring. At the same time, it is essential that only a single node out of all available nodes in the region should forward the data to the sink node. Otherwise, there would be redundancy in the data that wastes channel and energy resources. Appropriate data aggregation mechanisms should be developed for forwarding the required information to the sink node.

The next important aspect is the availability of sink node. The architecture should consider forwarding to multiple sink nodes simultaneously to protect data against sink node failures. It is possible that the communication link to a particular sink node for certain duration is highly noisy because of some nearby activity or the link is in deep fade due to surrounding weather conditions. Also the sink node storage capacity must be taken into consideration. It is very common now that the number of sensor devices are increasing enormously and each sensor device is pumping huge amount of data into the

network. For example consider the case of connected cars. The sensors in a car will be sending gigabytes of data during every hour of its running to the edge network. The edge network with destined sink nodes must have adequate capacity to collect data from the oncoming cars. Also due to high mobility, a particular car will be changing its sink node destination along this path of travel. So, the kind of application plays a key role in coming up with the right kind of architecture. All such varied degree of requirements must be considered in coming up with a single unified architecture.

Besides simple storage, are the sink nodes able to process and manipulate data appropriate for analysis? A better scalable architecture may or may not allow this. It is also possible that the sink nodes are not better equipped with tools and resources for processing the data. They are better equipped to efficiently store and simply forward to the right kind of servers in the core network. For high availability of data and that too in realtime is making the service providers push data to edge networks by placing servers close to the routers in these edge networks that are providing network access to end users. Such kind of architecture will profoundly improve latency issues in the network.

B. Cellular Networks

Cellular networks have emerged predominantly as single hop infrastructure based service networks to initially provide voice communication. Their main concern is reliable connectivity along with mobility. With the proliferation of wireless devices such as smartphones, tablets, and laptops, wireless communication have evolved into multimedia provisioning networks [3]. Users are now accessing videos, mobile TV, voice over IP, multimedia streaming, video chatting, e-commerce, banking, and many killer applications in social networking. The demand for higher bandwidth is therefore constantly raising. People are relying more on their smartphones and other devices for monitoring their homes and for other remote operations. People are spending more time online now than before and this trend is growing tremendously. New users are constantly added to the networks and so as new bandwidth intensive applications. Data movement in these networks is humongous and to keep with such growth HetNets are evolving [4]. HetNets have multi-sized cells such as macro, micro, pico, and femto cells. This is mainly because, smaller cells can provide higher bandwidths and reliable connections in highly dense locations where there is low mobility. Macro and micro cells on the other hand provide overlay coverage as an umbrella network for mobile users. Large cells reduce the number of handovers and as a result are effective in reducing the call drops.

Now the new requirement along with connectivity and mobility is bandwidth and affordability. The demand for data bandwidth is growing but the revenues generation for service providers is not growing proportionally. The network costs are constantly raising with newly adopted technologies, and if this trend is not changed, the upcoming 5G networks will be prohibitively expensive to do business for service providers. So, the prime concern of the next upcoming 5G network architecture is affordability of the new technology, how this new technology can be rolled out to the common users. The major portion of this expenditure is the energy required to keep the networks up and running and the operations and maintenance of the networks. New network consolidation models are required to reduce the operations cost and energy bills. The networks have to be self organizing to eliminate expensive manual interventions [5]. The new considerations of service providers is to reduce the energy consumption of their base stations [6].

Deploying smaller cells like pico and femto cells will profoundly reduce energy consumption in their base stations [7], [8]. For instance, femto cells can be user deployed which eliminates the burden on service providers to power and maintain them [9], [10]. It is expected that these femto base stations would be plug and play devices that are easy to install and operate. However this puts a much greater onus on service providers to synchronize transmissions between various femto base stations and other nearby pico, micro, and macro cells. Interference is a major issue between adjacent cells which are sometimes overlapping in coverage regions. The networks become highly complex as resource block scheduling MAC protocols and interference mitigation schemes have to work under very tight constraints.

While the new 5G architecture has to take all these factors into account, one more important factor that is our main focus is the generation and handling of Big Data in these networks. It is expected that data has to be available to users in realtime while they are moving between locations. Networks have to ensure that the data stored and retrieved is less expensive and less complicated. The data manipulation for analysis requires suitable architecture for display of results to right users within the right time frame. It is important to know that user devices such as smartphones, tablets, and laptops are the sources of data. Millions of users are always on the service providers network constantly generating data such as location information, services accessed, and other critical data. For instance, a safety application can use such data to provide rescue operations and alert the nearest fire and police department for immediate help.

Another major instance from where data is generated is the large and growing number of social networks. Data is shared between users in form of pictures, videos, files, information on certain activities, concerts, shows, and fashion. Like minded people will be discussing on a particular topic may it be of social nature, hobby, technology, or scientific. Conversations from all these sort of activities generate data on the network. Smart homes and smart cities generate data every single instant of time. Telemedicine, health monitoring, and reporting applications are gaining a lot more importance in old age people preferring to lead independent life. On the other hand, the news is constantly shared among people in communities on various political, technical, health, and entertainment issues. A unified architecture should also accommodate various technologies into a solution and promote applications working on cross platforms and multiple technologies simultaneously.

C. RFID Networks

Radio frequency identification (RFID) tags [11] are attached to every imaginable thing that is monitored such as any inventory, commercial goods and merchandize, home and business equipment, industrial production cycle for automation, citywide as well as countrywide infrastructure including

buildings and roadways, domestic and wild life animals, bird species migration habits, natural habitat, old age people and patients, etc., to name a few. RFID tags are sleeker than sensor devices and are less complex. They only store certain amount of data which will be read by a RFID reader. Sensor devices on the other hand have some sort of sensing functionality, which they use to produce data and forward it to their sink nodes. Sensor nodes can initiate communication with their sink nodes and have a MAC module built in them whereas RFID tags are not equipped with such MAC module. MAC intelligence in RFID networks is incorporated only in RFID readers, which read data from RFID tags and forward them to gateway node attached to central network [12]. These RFID readers can be sensors placed in ad hoc manner in between the gateway on the wired network and RFID tags attached to the monitored devices. In this manner both RFID networks and sensor networks work together to complete the required job [13].

RFID tags are either active or passive. Active RFID tags are considerably bulky as they are equipped with battery source and protective gear for the onboard battery. Passive RFID tags on the other hand do not have any battery onboard as they harvest power from the electromagnetic energy of the RFID reader communication. The main challenge with these RFID networks is the communication range. Though active RFID tags can communicate with longer range than passive RFID tags, still the communication range is only around 5 to 10 meters for active tags. Passive tags communicate only up to 1 to 2 meters. Further the amount of data they can store is very low due to size limitations. For instance the tags attached to many bird species should be less than 1 gram, which is less that 5% of the body weight as recommended by a standard norm. Further the data rate in these RFID is very low and is often one way communication, from the tags to the reader. However it is possible to have two-way communication in some active tags, but will increase the size and complexity. This is also because, the two-way communication may require a simple onboard MAC module on these active RFID tags. Another drawback with active RFID tags is that their lifetime is limited unlike passive RFID tags (which is unlimited).

RFID tags are also suitable for near field communications (NFC). These tags can be used in credit cards for transactions in stores, in passports to scan information on identity in airports, patient identification in hospitals, and many such applications. Though the data that is generated by each individual RFID tag is very less, the total data of collective millions of such tags in a certain location is huge. The data is very critical in offering services to the end users. The authenticity of the transactions is very critical as people rely on the technology in their daily lives. The system of RFID networks has to be reliable as business transactions are performed. Secondary nature data and analysis for research and study purposes for improving the business performance and any other such tasks is also highly essential. A proper combination of RFID networks and sensor networks can be designed for efficiently conducting daily business and operations. The unified network architecture has to consider all these factors in designing a robust monitoring system. It is particularly important to know that the information is not only used by the main business owners but also by general public over the Internet. Take for instance a certain consumer is going online on the Internet to look for available stock in a particular Walmart or Canadian Tire store. The data available on the stores website can be realtime data acquired from the the RFID tags on the available inventory.

D. Device-to-Device (D2D) Networks

A major challenge for the emerging 5G networks is to accommodate tremendously huge number of users, increased by many folds compared to the current number of users. The current number of users in cellular networks is already significantly high. The density of users is going to increase exponentially, which makes it very difficult for the base stations to provide service to all these users. Keeping this in mind, the service providers have already decided to reduce the cell sizes leading to HetNets. But the problem with handling these increased number of smaller cells leads to complex signalling. Further, the traffic has to be passed through the base stations of this cellular networks. Given the increased amount of social activity and proximal communications for applications involving photo sharing, games, chatting, etc., the traffic puts an enormous burden on the cellular network. It is possible to offload some of the traffic from cellular network by initiating direct communication between the parties that are proximally located and involved in some sort of communication. Such cases of proximal communications are predicted to be significantly high. It is, therefore, beneficial to develop a robust technological solution in this direction for offloading cellular network traffic in case of proximal communications. Keeping this in mind device-to-device (D2D) networks that provide direct communication between devices participating in communications is being developed [14], [15].

There are many benefits in D2D network communications because of their close physical proximity [16]. Better channel conditions prevail when devices are much closer. Higher throughput is possible under low signal to interference noise (SINR) ratios, leading to the possibility of having energy efficient communications [17]. D2D networks can be monitored by cellular base station through out the session by controlling it from initiation to termination [18]. This is defined as network controlled D2D communications. Network controlled D2D communications can be robust due to continuous control of the cellular base station. The advantage with this kind of D2D communications is that, they will not interfere with cellular communications or at least the observed interference will be low. Highly secure communications are possible through this scenario. It is also advantageous when devices are moving frequently. During high mobility, D2D sessions can switch to their communications through base station almost instantaneously and more seamlessly. The disadvantage of this kind of scenario is however that though data is transferred directly between communicating devices without passing through the cellular base station, still the control channel for these sessions is maintained through this base station. Due to this reason, not many simultaneous D2D network communications can be initiated within the cell region of a particular base station. This scenario is not highly scalable and not appropriate for dense D2D networks.

A more scalable solution for low mobility scenarios is partial assistance of base station during D2D communication initiation. Through this approach, it is possible to authenticate

the parties that are willing to participate in D2D communication. Since there is no further involvement of Cellular base station, no control channel is maintained for the ongoing D2D communication sessions which may last for very long durations. The resources of the cellular base stations are not tied up with the current ongoing, already established D2D network communications. This scenario is better scalable compared to the previous one and is also secure. As mentioned, this is a very good solution for low mobility situations where the devices are not expected to move from their positions. In practical situations, such conditions exist on a more frequent basis. Since the cellular base station is involved in setting up the D2D communication sessions, it can control and lower inference to the cellular communication network. This scenario can be regarded as network assisted D2D communications [19], [20].

It is possible to have autonomous operations for D2D communications [21]. These communications are established, maintained, and terminated without the need of a cellular base station. This is called assistance-free D2D communication, where cellular base stations are relieved from operations of D2D communications. Since cellular base stations are not involved in assistance-free D2D communications, there can be many D2D sessions, which is particularly suitable for dense environments. However, there is a growing concern of interference between D2D communications and cellular communications. There is no central authority controlling interference in the entire cell area. To overcome interference issues, D2D networks are using cognitive radio (CR) technology. It is possible to have either underlay or overlay CR communications in licensed frequency spectrum without interrupting primary users in cellular communications [22], [23]. D2D network users will be considered as secondary users to spatiotemporally available spectral resources. The secondary users ensure that their communications do not interfere with the primary users. They ensure that the total interference from their communications is below the tolerable level for primary users. In order to completely eliminate interference in the licensed bands used by cellular communications, D2D communications can use unlicensed bands. However, quality of service (QoS) in unlicensed bands is not guaranteed. As a result D2D communications will face more interference. For this reason, millimeter waves are being explored for D2D communications in upcoming 5G networks [24].

Due to varied degree of involvement of cellular base stations in the above three different scenarios, it best to leave it to the end users requesting the kind of service for their D2D network communications. Depending on the required QoS for D2D communications, based on their degree of mobility, security level requirement, and device capabilities, there would be an option to choose network controlled, network assisted, or assistance-free D2D service. The network architecture should embrace all these three types of service configurations and maintain statistics of the entire data communicated along with the level of data management and analysis required for the end Big Data applications.

II. Big Data Management Issues

Having glanced at various types of networks available around us that specialize in their services to diverse users, its time to move on to a more important issues surrounding the data generated in these networks. All of us will be interested in the answers to the following questions:

- What to do with the data generated in these networks on daily basis?
- Do we simply discard data or store it for later retrieval?
- What are the benefits of storing the data?
- How important is the data?
- Is the information derived from the data reliable to take action on it?
- Who will access the data and what are its implications?
- Who does it benefit?
- Are you comfortable in sharing your data and what are the consequences?
- How do you guarantee privacy and confidentiality?
- What kind of technology is needed for the job and how you compare with the current available technology?
- Is the available technology scalable according to our growing needs?
- Is the technology affordable and energy efficient to meet green objectives?

These are some very general questions pertaining to the data that is being produced daily by user activities in cities, countries, and around the planet. In addition to the surveillance activities, much information can be gathered from the daily generated data to know the preferences of users, their requirements, and problems faced. The data can be used to systematically tackle issues that are faced by general public in a more comprehensive and economical manner. Data will be useful while making meaningful analysis to improve technology and services for users. Social networks provide most of this information [25], [26], [27]. For instance, traffic pattern on a particular highway can be known from using communications infrastructure [28], [29]. From this information, the province can plan gas stations, motels, and other services needed during traveling. Expansion of highways and other planning activities need information on the kind of activity on the highway. Emergency response teams need timely data to deliver their services within time.

While taking appropriate measures in preserving sensitive private information from public disclose, it is highly beneficial to share data with right institutions and companies to enhance services and technology. Data sharing becomes vital during catastrophic situations such as an earthquake, fire, political and social disturbance, road accidents, construction, traffic congestion, and when any other unusual phenomena happen. Proper use of data will improve living and save lives during disasters [30]. People will be updated with knowledge and information upon access to required data. There are many other benefits such as gaining insights and relations to the surrounding activities related to human, nature, and technological interactions. It is possible to gain new grounds, and

avoid disasters before hand with the available data gathered from various sources.

To gain any of such new insights, huge unstructured data from many sources is needed. This huge data is available from surrounding activities and needs powerful computing platform and huge but compact storage space [31]. The technology should be scalable with the growing size of data and also be economical. Along with being affordable, the technology should be green using as little energy as possible. The technology should provide results in realtime or within the required time constraints. The results have to be reliable and verifiable. The rest of the paper will look into the technological requirements to handle data and in particular answer the last three of the questions listed above.

A network architecture should address all these issues such as scalability, energy efficiency, time and space constraints [32]. At the same time, the architecture has to deal with heterogeneous networks and the various kinds of devices connected to these networks. The data generated on these devices are from various applications and services leading to the handling of highly unstructured data. The devices have varied capabilities, various network topologies exist, different protocols, and conflicting objectives. All these factors have to be taken into consideration for handling Big Data generated from IoT devices.

A. IP Device Capabilities

The main characteristic of IoT devices is that they have varied device capabilities. These capabilities are in terms of computational power, storage capacity, energy constraints, and network connectivity. For instance, sensor devices have very limited battery capacity and most of the time it is hard and almost impossible to recharge the batteries. In such cases, the main focus will be to increase the lifetime of the sensor devices as much as possible. So they use the concept of duty cycle in their transmission operations. This is because, most of the battery energy is consumed by transmission and reception activities. In a duty cycle, there will be a sleep period and wake-up period. During sleep period, the devices turn off their transceivers and enter sleep mode. They wake up after the end of their sleep duration. They can transmit and receive only during the wake up period. It is typical to have 50% or less duty cycle to ensure longer sleep periods to enhance battery life.

Sensor devices constantly gather environmental data and need onboard storage to store data and transmit during the wake up period. Since these devices are not equipped with large storage capacity, the duty cycle should be adjusted accordingly. If the sensing activity is not very intense or frequent, then the devices can have very large sleep durations. The MAC operations have to be turned optimally to the sensing activity carried out by the devices. The MAC layer, therefore, has to ensure that battery life is elongated along with efficient transmission of the sensing data to its sink node. The network architecture has to consider these aspects and may need to provide multi-hop communications to conserve battery life in sensor devices. It is important to synchronize transmissions among sensor devices such that the relevant communicating partners are awake at the same time and go to sleep mode in a synchronized fashion.

Sensor devices are not meant for complex processing of the data onboard. So, they do not have processors on board. Most of the time they are transducers that convert one form of energy into another and send the raw data over the sensing activity to their central processing and storage sink nodes which are considered more powerful than these sensor devices. The sink nodes will preprocess the data and later forward the central network for complex analysis and permanent storage of the data in appropriate form.

Cellular network devices are more powerful. They are mainly smartphones, tablets, and PDAs that transmit data, voice, and multimedia containing audio and video communications. They have powerful onboard processors but rely heavily on central base station for complex network scheduling operations. Communications in these networks are two-way, although these communications are asymmetrical in uplink and downlink bandwidth occupancies. The download communications from central base station to the mobile terminals are bandwidth intensive compared to the uplink communications going from mobile terminals to the central base station. This is particularly true with data communications where the mobile terminal will be constantly downloading data from the central servers on the Infrastructure based network through the base station. Compared to sensor devices, these devices are easily rechargeable. However, energy considerations are still vital due to limited battery capacity. These devices also use some sort of sleep cycle to conserve energy in their batteries. MAC operations in these devices are more power hungry due to advanced applications that are built for these type of devices.

RFID tags are passive devices which harvest energy from RF communications of their RFID readers. They have no onboard processing and are suitable for very short range communications. They may not have MAC module and are reactive to the communications from the RFID reader. The communication will be mainly one way from RFID tags to the central reader. The RFID readers will be connected to the main central server in the Infrastructure based network. The RFID readers are simply bridges from the RFID networks to the central core data networks. Data will be processed and stored in the central core data networks.

Multihop D2D communications are being incorporated in cellular networks to conserve energy in mobile devices. When devices are in close proximity with each other, direct communications are possible between these devices without any need of services from the central base station. Thus central base station can be relieved from providing services to such directly possible communications between close vicinity mobile terminals. D2D communication networks are being developed to take advantage of such network communications. The devices will be able to setup, manage, and tear down communication sessions among them, thus offloading some of the traffic from the central base station. D2D devices have to be more complex to avoid interference with the cellular communications that are present overlapping with the regions of D2D communications.

B. Network Considerations

Network topology has to be robust to maintain both physical and logical links all the time. Network redundancy must

be maintained to overcome intermittent failures, which are commonly known as backup and restoration mechanisms. The network redundancy can also be used to provide load balancing features into the network. For instance, it is critical in sensor networks to have such type of network redundancy. If a sensor node failure occurs due to adverse environmental conditions or due to battery failure, it may bring the whole network down depending on its location in the network. Such bottlenecks have to be eliminated by maintaining alternative routes in the network. It is very challenging to have such mechanism particularly when the network is setup in ad hoc manner. Generally, most of the sensor networks are ad hoc networks. Take for instance the case of monitoring a forest area. The sensor nodes are sprayed using a means of air transport such as a helicopter. As the sensor nodes are sprayed over the area, their locations in the overall sensor network are completely random. To avoid frequent node failures, it is best to increase the density of the nodes to have both backup paths and provide load balancing in the network.

The nature of communication in sensor networks is quite the opposite to that of the communications in cellular networks. The communications in sensor networks is always from many leaf sensor nodes to the single central sink node. Whereas in cellular networks, most of the data is downloaded by the mobile terminals from the central base station. Due to this stark contrast, the design of MAC protocols and the network topology design will be completely different. Sensor nodes are seldom mobile. However, there are upcoming sensor network applications that involve high mobility of sensor nodes, such as sensors in connected cars. Varied mobility ranges have to be taken into consideration depending on the type of application at hand. However, in both sensor and cellular networks, critical data is maintained in the central servers which receive data from the base stations or sink nodes. The network topology should ensure that these nodes do not become single point of failures. Communication has to be maintained always through more than one sink node in sensor networks to overcome network and link failures. When there is mobility, sensor nodes need proper handoff mechanism to relay their communications with the next neighboring sink node along the path of travel.

Besides mobility, there has to be enough bandwidth for data to be transferred over wireless media to servers located in the central infrastructure based network connected to the Internet. This is how the data is made available for global access. The time delay is highly critical for realtime applications. The data that is fetched by the mobile terminals should be available readily and on demand. The accessed data will be critical for maintaining the safety and proper operation of the devices attached to the Internet through onboard sensors. Take for instance the highway patrol team using technology attached to the sensor networks. The data carried over these sensor networks is critical for the operations of the highway patrol team to provide emergency services, redirect traffic, and provide warnings for the safety of the travelers on the road. Network congestion should be avoided for proper operation of such network. Adequate network bandwidth has to be provisioned. In addition to the bandwidth, network security is very important. The transmissions have to be highly secure to avoid eavesdropping on the critical communications. Security is needed for protecting the identity of the individuals and any unauthorized access to the personal information.

Providing high bandwidth over wireless media is highly challenging. This is due to the fact that the available bandwidth over wireless channels is very low compared to the existing wired media such as fiber optics. Wireless links are also challenged with link fading conditions due to multipath transmission, obstructions, weather changes and many other varying surrounding conditions. Because of this kind of fading phenomenon, constant channel bandwidth cannot be guaranteed for wireless links. Wireless links are also subjected to interference from surrounding activities. Due to all these reasons, the signal to interference noise (SINR) ratio of wireless links constantly varies. Mobility creates Doppler effect and the distance between the transmitter and the receiver is not always the same. As a result, communication links have to be constantly changed between the mobile terminals and the closely available base station. This is commonly known as handoff mechanism. The handoff should be seamless to the ongoing transmissions to maintain good quality of service (QoS) and quality of experience (QoE).

One way to increase the bandwidth is to reduce distance between transmitters and receivers involved in wireless communication. Higher SINR is possible over shorter distances, which can be useful for adopting high modulation and coding rates (MCR) to improve spectral efficiency measured in terms of (Bits/s)/Hz. High MCR will increase the transmission data rate on the wireless channel. Further, reducing the link distance between the transmitter and receiver will improve line-of-sight communication between the two transceivers involved in wireless communication. The next advantage is, it is possible to use higher frequency spectrum over shorter distances. Higher frequency spectrum provides higher bandwidths that can be useful for multimedia transmissions. The link distances have to be shorter because fading is heavy over higher frequency ranges compared to the lower frequency range. Higher frequency ranges need line of sight communications. It is possible to reduce SINR of the transmissions to conserve energy that is critical to extending the battery life.

Because of this reason, cellular communications are adopting smaller size cells to improve data rates. Smaller cells will improve frequency reuse over space and time. More number of users can be supported with many smaller cells instead of a single large cell. This is a dire necessity for today's communication requirements as more users are being added to these cellular networks every day. The number of users is not only increasing but also all these users are demanding higher bandwidths at higher mobility speeds. As a result the networks have become complex in terms of signalling. Maintaining many smaller cells increases scheduling complexity because of high level of cross-tier and inter-tier interferences due to transmissions between the cell base stations and its associated users. High mobility and smaller cells means increased number of handoffs for mobile users. This is an additional load on the cellular infrastructure. Traditional cellular communication requires all transmissions passing through the base station. This is required to maintain adequate quality of service and better control of the ongoing communication sessions.

In order to maintain short link distances, and improve QoE, cellular networks are offloading some of the traffic to direct D2D communications in cases where the transceivers are in

close proximity of each other. These D2D transmissions need not pass through the base station thus relieving the cellular network from some of this kind of traffic. Multihop D2D communications are not only energy efficient but also are required for the large data transfers between closely located users. Such kind of communication scenarios are quite common where users that are close by are sharing photos, videos and other personal data files with each other. This type of network topology will substantially enhance the spectral efficiency over space and time. The network topology empowers the mobile terminals to take routing decisions for multihop transmissions and relaying to the base station to conserve energy and improve communication range. This type of communication can also be used to improve QoE for the users located at the cell edges.

C. Energy and Cost Considerations

Trillions of IoT devices around the world, which are connected to the Internet, consume a significant amount of energy. Higher data rates and increased processing powers make these devices power hungry, consuming more power. As the number of these devices continue to increase exponentially, there is a risk of consuming significant portion of the global energy resources. New initiatives for greener communications are taken to reduce global energy consumption footprint. In existing cellular mobile networks, significant portion of energy is consumed by base stations. In addition to that, backend servers providing services to users and application servers in data centers also consume a lot of energy. Until now, the power saving considerations were made for mobile terminals and sensor devices that are difficult to recharge. While, this is very important and required for satisfying operating constraints, it is now being looked in for saving energy in the network infrastructure as well.

The other important consideration is the involved cost. Lowering infrastructure and operational cost are essential for businesses in communication sector to remain profitable. Communication networks are constantly evolving, and new technologies are developed each year. It is very difficult for the business enterprizes to embrace all these new powerful technologies and remain updated for offering better services. Higher bandwidth offerings and the addition of new value added services to its users is not leading to a proportional increase in the revenue generated. Investment costs are skyrocketing, and profit margins are becoming tight. Now the point has come to revolutionize the adopted technology with new architectures to substantially lower investment and operational costs. This also includes changing the business model and ownership of the network infrastructure.

There is a huge cost to manage Big Data across different networks. As the size of the data grows, the cost for handling the data also grows. This growth in cost has to be reduced by innovative strategies that involve optimizing the utilization of hardware resources. Moving the data around the network involves communication costs. However, keeping the data available at fingertips needs a constant movement of data around different parts of the network and keep the frequently accessed chunk of data close to the users to avoid excessive delays. While redundancy in storing data is a favorable feature against network failures, it increases the cost of storage per unit of data. Therefore judicious strategies must be implanted to reduce the cost of storage. An optimized strategy should be adopted that balances storage cost and availability. It must not be forgotten that more required storage leads to higher energy consumption.

III. Architecture for Heterogeneous devices and network types

Based on all the considerations that are discussed, that includes the different network types and heterogeneous devices attached to these networks, a single unified architecture for central management is an ambitious goal to achieve. However, that is the trend which is being adopted for reducing the operational costs. The ease of network management is the prime objective to reduce the burden of hiring and retaining network specialists. Network automation minimizes the delays that requires human intervention, and improve service quality. With automation, the network resources are efficiently managed. It is possible to operate networks in their optimized regions, leading to achieve better price versus performance results. More users can be accommodated and managed with less effort.

Starting with this simple objective of reducing operational costs, the solutions that were seem feasible to deploy have more potential benefits. Those benefits are reduced amounts of complex and specialized hardware through replacement with software based systems. Software systems can be better programmed according to the needed requirements, which configuring specialized hardware will do exactly the same operation. A solution in hardware is expensive than performing the same job through software programming on a general purpose device. It is easy to test new things and reprogram software systems. This philosophy is being adopted by business enterprizes comprising network operators. They are replacing specialized network hardware equipment with general COTS based servers. The entire network functionality is realized in software implementation on these COTS servers. The other derived benefit of COTS servers is that they are more power efficient and consume less energy than the specialized network hardware.

The next highly preferred feature is a network solution that is scalable and easy to upgrade. Hardware solutions have many limitations to the scalability. Just adding new specialized servers to the rack is not always possible. After a certain stage, the entire hardware need to be replaced causing the solution to be expensive. The network operator will have to decide to move to the new solution or keep the old one due to space limitations in the server room. It is hard to revert back to the previous working scenario in hardware solutions if there is any problem with the new technology. Changes to the hardware equipment will cause network downtime. All these can be easily addressed in software configurable systems. New COTS servers can be added to the server farm as need arise to increase the computational power and storage space.

New solutions need to handle and manage Big Data with ease. Further, these new network solutions have to be cheap to sore and process mammoth data. The costs and energy consumption can easily spiral without the right technology and the right solution. Any solution to keep Big Data at bay will be in large scale. This has to be kept in mind. The solution needs

hundreds of thousands of servers for processing and storing Big Data. Where to place these number of servers and how to power them and manage the heat generated by them is a big issue. Besides huge sized server farms, the management complexity should be minimal. Typically these data networks will be clusters of distributed server farms connected through high bandwidth links such as fiber optics. Latency is an issue when these server farms are far located. The data usage should be predicted and kept at optimal locations across the globe for instant access.

IV. Big Data delivery and Information Processing

The main function of future IoT network infrastructure is to be able to deliver Big Data. Data from bits and pieces are amalgamated to form huge unstructured data providing many insightful ways of information for human needs and development. It is not just the infrastructure that is important, it is also important to know what we do with the existing infrastructure. How we build powerful applications that manage and harvest much needed information for our daily business. IoT infrastructure is the backbone for Big Data generation. On demand resources have to be provisioned by the network infrastructure to deliver Big Data to the right places at the right time. Many challenges remain in handling Big Data and one of this is how to efficiently deliver data. Data communication is not cheap as it requires tremendous resources and requires proper network setup and maintenance. Huge energy resources have to be allocated for data transmission. Servers that act as network routers consume power in the order of Mega watts per each day. It is, therefore, essential to know the power consumption per unit of data transfer.

Along with data transfer, is the storage and processing resources for the Big Data. New technologies are needed to reduce the storage cost by increasing the memory size per unit area. High density storage equipment will, in turn, reduce the amount of building space needed for the storage servers. The running costs for storing data is the power consumption of these servers, the airconditioning of the premises, and other wear and tear costs that are incurred on a daily basis. In addition to this is the manpower needed for maintaining the server farm. The computational power of today's available GPUs is high but consume a tremendous amount of power. Efficient data manipulation softwares are needed to intelligently compute data and derive results. These intelligent softwares must use lesser CPU cycles to achieve energy savings. Data analysis and information processing are critical to handle and maintain data. During computation, analysis, and storage, communication protocols in the network play a vital role.

V. Conclusion

Different kinds of communicating devices that are made for completely different objectives are communicating with each other over the IoT infrastructure. Different types of networks exist today giving connectivity to heterogenous devices which are both mobile and static. Trillions of such devices exist today giving rise to Big Data. New applications are being developed to harness information from proper handling of Big Data. At the same time network operators are faced with the dilemma whether or not to embrace new technologies because of tighter profit margins. Technology is changing and new innovations are changing the way network communications take place, and that too at a fast pace. For all these reasons, new architecture of IoT is needed to address the concerns of profitability, better services, technology adaption, user growth, and energy conservation. For this reason, we have thoroughly discussed the various issues involved in building an architecture for IoT networks that requires Big Data Management.

References

[1] A. Mahapatro and P. M. Khilar, "Fault diagnosis in wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2000–2026, Fourth Quarter 2013.

[2] I. Demirkol, C. Ersoy, and F. Alagöz, "Mac protocols for wireless sensor networks: A survey," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 115–121, April 2006.

[3] K. Hamad, S. Primak, M. Kalil, and A. Shami, "Qos-aware energy-efficient downlink predictive scheduler for ofdma-based cellular devices," *to appear in IEEE Transactions on Vehicular Technologies*, 2016.

[4] H. Boostanimehr and V. K. Bhargava, "Unified and distributed qos-driven cell association algorithms in heterogeneous networks," *IEEE Transactions On Wireless Communications*, vol. 14, no. 3, pp. 1650–1662, March 2015.

[5] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 336–361, First Quarter 2013.

[6] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 4, pp. 524–540, Fourth Quarter 2011.

[7] M. Kalil, A. Shami, A. Al-Dweik, and S. Muhaidat, "Low-complexity power-efficient schedulers for lte uplink with delay-sensitive traffic," *IEEE Transactions on Vehicular Technologies*, vol. 64, no. 10, pp. 4551–4564, October 2015.

[8] M. Kalil, A. Shami, and A. Al-Dweik, "Qos-aware power-efficient scheduler for lte uplink," *IEEE Transactions on Mobile Computing*, vol. 14, no. 8, pp. 1672–1685, August 2015.

[9] K. A. Meerja, P.-H. Ho, and B. Wu, "A novel approach for co-channel interference mitigation in femtocell networks," in *proceedings of IEEE Globecom 2011*, December 2011.

[10] K. A. Meerja, P.-H. Ho, B. Wu, and H.-F. Yu, "Media access protocol for a coexisting cognitive femtocell network," *Computer Networks*, vol. 57, no. 15, pp. 2961–2975, October 2013.

[11] Z. Yang, T. Ning, and H. Wu, "Distributed data query in intermittently connected passive rfid networks," *IEEE Transactions On Parallel And Distributed Systems*, vol. 24, no. 10, pp. 1972–1982, October 2013.

[12] M. Shahzad and A. X. Liu, "Fast and accurate estimation of rfid tags," *IEEE/ACM Transactions On Networking*, vol. 23, no. 1, pp. 241–254, February 2015.

[13] J. Cho, Y. Shim, T. Kwon, and Y. Choi, "Sarif: A novel framework for integrating wireless sensor and rfid networks," *IEEE Wireless Communications*, vol. 14, no. 6, pp. 50–56, December 2007.

[14] L. Wang, L. Liu, X. Cao, X. Tian, and Y. Cheng, "Sociality-aware resource allocation for device-to-device communications in cellular networks," *IET Communications*, vol. 9, no. 3, pp. 342–349, February 2015.

[15] S. Sun, Q. Gao, W. Chen, R. Zhao, and Y. Peng, "Recent progress of long-term evolution device-to-device in third-generation partnership project standardisation," *IET Communications*, vol. 9, no. 3, pp. 412–420, February 2015.

[16] K. W. Choi and Z. Han, "Device-to-device discovery for proximity-based service in lte-advanced system," *IEEE Journal On Selected Areas In Communications*, vol. 33, no. 1, pp. 55–66, January 2015.

[17] G. Wu, C. Yang, S. Li, and G. Y. Li, "Recent advances in energy-efficient networks and their application in 5g systems," *IEEE Wireless Communications*, vol. 22, no. 2, pp. 145–151, April 2015.

[18] A. Moubayed, A. Shami, and H. Lutfiyya, "Wireless resource virtualization with device-to-device communication underlaying lte network," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 734–740, December 2015.

[19] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto d2d links in unlicensed bands," *IEEE Journal On Selected Areas In Communications*, vol. 33, no. 1, pp. 67–80, January 2015.

[20] Y. Ni, X. Wang, S. Jin, K.-K. Wong, H. Zhu, and N. Zhang, "Outage probability of device-to-device communication assisted by one-way amplify-and-forward relaying," *IET Communications*, vol. 9, no. 2, pp. 271–282, January 2015.

[21] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Transactions On Communications*, vol. 63, no. 2, pp. 441–454, February 2015.

[22] M. G. Khoshkholgh, Y. Zhang, K.-C. Chen, K. G. Shin, and S. Gjessing, "Connectivity of cognitive device-to-device communications underlying cellular networks," *IEEE Journal On Selected Areas In Communications*, vol. 33, no. 1, pp. 81–99, January 2015.

[23] Z. Zhou, M. Dong, K. Ota, R. Shi, Z. Liu, and T. Sato, "Game-theoretic approach to energy-efficient resource allocation in device-to-device underlay communications," *IET Communications*, vol. 9, no. 3, pp. 375–385, February 2015.

[24] J. Qiao, X. S. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5g cellular networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 209–215, January 2015.

[25] M. Jammal, T. Singh, A. Shami, R. Asal, and Y. Li, "Software-defined networking: State of the art and research challenges," *Computer Networks*, vol. 72, no. 0, pp. 74 – 98, October 2014.

[26] M. Jammal, A. Kanso, and A. Shami, "Chase: Component high availability-aware scheduler in cloud computing environment," in *Cloud Computing (CLOUD), IEEE 8th International Conference on*, June-July 2015, pp. 477–484.

[27] ——, "High availability-aware optimization digest for applications deployment in cloud," in *IEEE International Conference on Communications (IEEE ICC)*, June 2015.

[28] H. Hawilo, A. Shami, M. Mirahmadi, and A. Asal, "Nfv: State of the art, challenges and implementation in next generation mobile networks (vepc)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, November-December 2014.

[29] H. Hawilo, A. Kanso, and A. Shami, "Towards an elasticity framework for legacy highly available applications in the cloud," in *Services (SERVICES), 2015 IEEE World Congress on*, June-July 2015, pp. 253–260.

[30] K. Alhazmi, M. A. Sharkh, and A. Shami, "Drawing the cloud map: Virtual network provisioning in distributed cloud computing data centers," *to appear in IEEE Systems Journal*, 2016.

[31] K. A. Meerja, A. Shami, and A. Refaey, "Hailing cloud empowered radio access networks," *IEEE Wireless Communications*, February 2015.

[32] M. A. Sharkh, A. Kanso, A. Shami, and P. Öhlén, "Building a cloud on earth: A study of cloud computing data center simulators," *Computer Networks*, vol. 108, pp. 78–96, October 2016.