TIME TO RETHINK
REAL-TIME BIG DATA ANALYTICS

Real-Time Big Data Analytics (RTBDA) has emerged as a new topic in big data discussions. The concepts underpinning RTBDA can be applied in a telecom context, but it does require a rethink on what is "real time" and what sources of information can support a RTBDA strategy. Fortunately, the technologies required to implement such a strategy are not only available, but in many cases already being used, just not as effectively or strategically as they could be.

Real-Time
Big Data Analytics
(RTBDA)

TIME TO RETHINK
REAL-TIME BIG DATA ANALYTICS

Real-Time Big Data Analytics (RTBDA) has emerged lately as a new term in big data discussions. It refers to one of the key aspects and value propositions of big data analytics, namely the ability to act either proactively or reactively in real time based on analysis of available information. This strategy is one of the cornerstones of many Internet companies, such as Amazon and Google.

These Over-The-Top (OTT) players are a source of inspiration, but also frustration to telecom carriers who need to come to grips with the increasing amounts of traffic that these companies are generating in telecom networks, with little or no revenue contribution. The first step is understanding why these companies are successful, which leads to an interest in how OTT players use big data analytics, and RTBDA in particular, to succeed.

In this paper, we will take a closer look at RTBDA and discuss it in the context of telecom networks. What we will discover is that the concepts underpinning RTBDA can be applied in a telecom context, but it does require a rethink on what is "real time" and what sources of information can support a RTBDA strategy. Fortunately, the technologies required to implement such a strategy are not only available, but in many cases already being used, just not as effectively or strategically as they possibly could be.

DEFINING RTBDA

In a very simplified way, one could say that big data analytics is composed of two parts that distinguish it from business intelligence or data warehousing and mining: distributed, parallel processing and the ability to act in real time.

One of the challenges that big data analytics addresses is the need to process large disparate data sets that normally cannot be accommodated by a single database or server. One of the solutions to address this is the use of distributed, parallel processing where large data sets are distributed

to multiple servers that each process a part of the data set in parallel. Big data analytics does not require a specific structure for the data, but can work with both structured and unstructured data. Using Hadoop with MapReduce is an example of such an approach and can be credited with being a driving force behind the current interest in big data.

There are now a multitude of complementary solutions to various aspects of distributed, parallel processing, including Cassandra, Impala, Hive, HBase, Storm and Spark, and many more. New tools are emerging every day to make it easier to program algorithms or analyze data. One of the driving principles of many of these solutions is time limitation. Solutions can be found for processing large amounts of data, but what is important in a big data perspective is that processing should be completed within a defined time frame. That time frame is now increasingly being associated with real time.

RTBDA is relatively new, but addresses the need to act proactively or reactively in real time. It is inspired by the capabilities of Internet content and services providers to understand what is happening, analyze the situation and take action in real time. For example, when new recommendations are displayed once you have made a purchase on a website.

For a better understanding of RTBDA, a good introduction is provided in "Real-Time Big Data Analytics: Emerging Architecture."[1] It provides a good overview of the technologies involved and the architecture of RTBDA solutions, as well as a five phase process for RTBDA implementation.

---

1 Mike Barlow, "Real-Time Big Data Analytics: Emerging Architecture," Sebastopol: O'Reilly Media, 2013.

DEFINING "REAL TIME" FOR TELECOM

In "Real-Time Big Data Analytics: Emerging Architecture," Mike Barlow asks the question, "How real is 'real time?'" The answer is: it depends. It depends on the context of what you are trying to achieve and the environment in which you are working. For some, seconds or microseconds are enough; for others, real time needs to be faster.

From a telecom point of view, this is an interesting question. It exposes a potential issue with current practices in telecom that needs to be addressed if carriers are to succeed in tackling the challenges that OTT traffic is posing. The fact is that the current acceptance of what is "real time" in telecom may no longer be sufficient.

In the past, telecom networks were based on connection-oriented technology and protocols (e.g. SONET/SDH). This meant that services and their supporting connections were planned, engineered and deployed in a static manner. Changes could only be applied centrally in a highly structured process. This meant that the network did not change very much from one minute –or even one hour – to the next.

In this environment, it was sufficient to gather information from the network at regular intervals to know what was happening. The protocols that were used were also rich in management information, so a great deal of insight could be gathered from just one protocol header.

In this kind of environment, "real time" can be defined in seconds or even minutes, which is why it is sufficient to collect Call Detail Records (CDRs) every 5 to 15 minutes to gain full insight.

The issue is that this is not the situation in telecom today. With the migration to LTE, telecom carriers have completed the transition to packet networks based on Ethernet and IP, which function in a completely different way compared to connection-oriented technologies and protocols.

First, the fundamental principle of IP networks is that the network takes care of itself. The network defines the path that traffic takes and reroutes that path depending on congestion and other conditions. This allows the network to react quickly to changes. The downside is that you cannot predict with certainty where traffic will be flowing. This is not made any easier by the fact that Ethernet and IP protocols, by design, do not contain the same level of management information overhead that connection-oriented protocols provide.

Packet networks are also bursty and dynamic in nature. They are designed to support multiple services consumed by multiple users sharing the same infrastructure. Over a long time period, it can look like the utilization of the network is quite low, but that is because traffic is transmitted in bursts, which can consume the entire bandwidth available. In such situations, the IP network is expected to react and make sure that this traffic is routed in a balanced way through the network. The bottom line is that changes can occur in the network from one IP packet or Ethernet frame to the next (see Figure 1).

When you consider that Ethernet frames in a 10 Gbps network can be transmitted with as little as 67 nanoseconds between each frame, then we begin to understand what "real time" means in a packet network. It is not minutes; it is not even seconds. It is nanoseconds.

This is the fundamental issue with how telecom network management and data analytics is being performed today. They both rely on CDRs, Event Detail Records (EDRs) and IP Detail Records (IPDRs) to understand what is happening in real time. But, this definition of "real time" is anchored in the paradigm of the past when polling every few minutes was enough. With packet networks today, "real time" means nanoseconds.

REAL-TIME DECISION MAKING

Just to be clear, using CDRs, EDRs and IPDRs for big data analytics is a good idea. It just depends on what you are trying to achieve. Big data analytics can be used for two broad categories of decision making:

• Better planning and optimization of services and networks based on trends and predictive analysis
• Real-time decision making

Using detail records for better planning and optimization along with other structured and unstructured data sources is appropriate and valuable. They are rich in information and useful trends and predictions can be generated based on this data. However, they will never provide a complete picture until they are complemented by real-time information from packet networks that can provide exact details on what happened, when.

Needless to say, detail records cannot be used for real-time decision making. They are only collected every 5 to 15 minutes, which is not compatible with our understanding of what real time should be in packet networks. For true real-

time decision making, it is necessary to continuously collect, store and analyze network information. To understand what is happening, all the relevant Ethernet frames and IP packets need to be examined in real time.

By capturing and storing network information in this manner, we not only enable the ability to analyze and act on this information in real time, but also provide a source of detailed, reliable information on what happened, when in the network that can complement other big data analytics activities.

IMPLEMENTING RTBDA IN TELECOM

In "Real-Time Big Data Analytics: Emerging Architecture" a good overview is provided of a potential layered architecture for implementation of big data analytics and a five phase approach to implementation. A number of technologies for performing analytics are detailed, so readers are encouraged to read this to understand how the upper layers of big data analytics can be implemented.

In this paper, we would like to focus on the real-time data collection layer. This layer can provide a constant stream of actionable information for decision making.
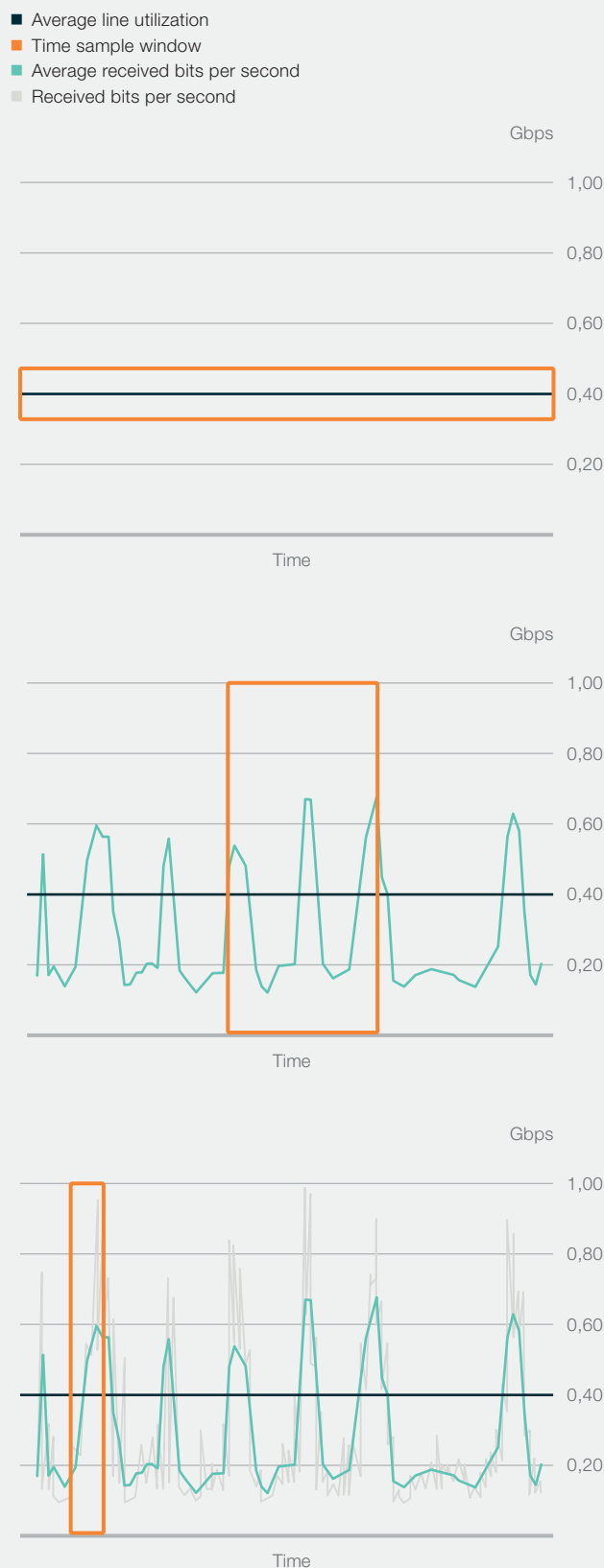
Both the TM Forum and the IP Network Monitoring for Quality of Service Intelligent Support (IPNQSIS) project, part of the European CELTIC-Plus program, have researched this need as part of their respective work on customer experience management. The conclusion from both projects was that probes and appliances are critical to providing reliable, real-time insight into what is happening in the network.

Probes are traditionally data collectors providing information to other management systems. Appliances, on the other hand, use the same technology, but also analyze the information and can store the information locally. Appliances are typically focused on a specific task, such as performance monitoring, test and measurement, or security, and are often seen as fulfilling that very specific role. But, probes and appliances can be used more strategically as sources of real-time data for big data analytics and as implementations of RTBDA strategies. The following provides a three step view of how such an infrastructure could be implemented.

DEPLOYMENT

The first step involves deployment of appliances for data collection. The key requirement here is that all the Ethernet frames and IP packets need to be captured, in real time, at line speed, with zero packet loss, no matter the conditions. This ensures that a reliable stream of information is being collected.

FIGURE 1
Utilization of Packet Networks

■ Average line utilization
■ Time sample window
■ Average received bits per second
■ Received bits per second

Combined Network Port Speed	Frames per second			Combined Network Port Speed	Time between packets		
	64B	300B	1518B		64B	300B	1518B
2 x 1G	3,0 M	0,8 M	0,2 M	2 x 1G	336 ns	1.280 ns	6.152 ns
2 x 10G	29,8 M	7,8 M	1,6 M	2 x 10G	34 ns	128 ns	615 ns
2 x 40G	119,0 M	31,3 M	6,5 M	2 x 40G	8 ns	32 ns	154 ns
2 x 100G	297,6 M	78,1 M	16,3 M	2 x 100G	3 ns	13 ns	62 ns

FIGURE 2
More Data, Less Time to React

It is also extremely important that each and every frame is given a unique time stamp so that an accurate timeline can be established, not only local to the appliance, but also across multiple appliances. The accuracy and precision of these time stamps must be in the range of nanoseconds. For example, with only 67 nanoseconds between Ethernet frames in a 10 Gbps network, the time stamp resolution must be better than 67 nanoseconds, otherwise two Ethernet frames would receive the same time stamp, making it difficult to distinguish which came first. In a 100 Gbps network, this reduces to 6.7 nanoseconds.

The combination of zero packet loss capture with nanosecond precision time-stamping ensures that we have a reliable, accurate stream of data analysis information.

STORAGE
The second step is storing this information in real time. Many appliances provide capture to disk, which allows real-time data to be stored directly to a local hard disk on the appliance. Alternatively, this data can be forwarded to a Storage Area Network (SAN) or other location. The stored data can be used to build a historical timeline of what has happened in the network with precise details. It is possible to recreate exactly what happened, as it happened, using this information.

For data analytics, this is a source of rich information. It can provide insight into usage and behavior trends. If the appliance has Deep Packet Inspection (DPI) capabilities, then usage of services, including OTT services can be tracked and analyzed to provide usage patterns with respect to time, location and type of device.

This information alone provides a valuable resource for network and service optimization. New, attractive services can be defined that match users' preferences. But, perhaps even more importantly, this information can be used to provide insight to OTT content service providers, so that carriers can offer compelling service offerings to these potential customers.

REAL-TIME DECISIONS
What is most exciting is the potential to use real-time data and stored data to enable real-time decision making, which is the third and final step. Historical information captured to disk can be used to develop a profile of expected behavior. When this is compared with the real-time information on what is going on in the network right now, it is possible to detect unexpected events or anomalies. These can be a security threat, performance degradation or an opportunity to offer a customer a package extension or a complementary service.

From a RTBDA perspective, this is very close to the types of abilities that OTT content and service providers have implemented. The ability to react in real time based on an understanding of what is happening right now, compared to what has happened in the past.

APPLIANCES: A STRATEGIC FOUNDATION FOR RTBDA
With this three step implementation, it can be seen that real-time information provided by probes and appliances can be used to implement RTBDA in telecom networks as well as complementing other sources of information in big data analytics for strategic planning.
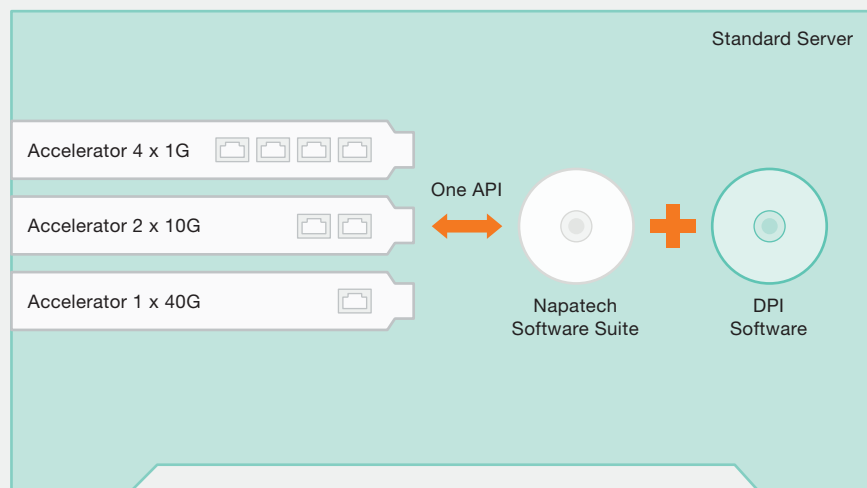
FIGURE 3
DPI Appliance

The technologies and products to implement this strategy are not only available, but are already being used, just not with this purpose in mind.

What is even more interesting is the fact that the vast majority of appliances being used today in enterprise, financial, government and telecom networks are based on commercial off-the-shelf server technology, which is compatible with carrier plans for the future. A cornerstone of both Software Defined Networks (SDN) and Network Function Virtualization (NFV) strategies is the use of off-the-shelf server hardware.

REAL-TIME TECHNOLOGY FROM NAPATECH

Napatech provides the world's largest vendors of appliances with the key technology and products that make it possible to implement appliances on standard server hardware. Napatech accelerators for network management and security applications assure zero packet loss and guaranteed delivery of analysis data, no matter the network conditions.

Napatech time-stamping is prepared for 100 Gbps networks with 4 nanosecond resolution. This feature allows precise measurement of jitter and delay in mobile networks, enabling not only traffic bandwidth analysis of OTT services, but also QoS analysis of these services.

To effectively separate and analyze each service used by a subscriber, it is necessary to extract information from the encapsulated IP header. Napatech accelerators also identify, filter and forward network flows based on layer 1 to 4

information, including the contents of IP-in-IP and GTP tunnels, along with other off-load features that ensure that the maximum amount of processing power is available to the analysis application running on the appliance.

We also work with the largest independent vendors of DPI software to provide an integrated network and service analysis solution. It is these capabilities that have made Napatech the undisputed global leader in this space and a de facto standard for appliance vendors.

TIME TO RETHINK RTBDA IN TELECOM

It is time to reconsider what "real time" means in modern telecom networks. It is also time to reconsider what sources are used for big data analytics. Telecom carriers must start using probe and appliance technology already in the network in a more strategic way to support RTBDA. Doing so will not only provide a better source of information for planning decisions, but will also open new opportunities to offer better services, not only to end users, but also to OTT service providers, which could finally address the issue of monetizing OTT traffic in telecom networks.

DISCOVER THE POWER OF NAPATECH

Napatech accelerators are designed to handle the maximum theoretical throughput of data for a given port speed. Napatech offers a range of accelerators supporting speeds from 10 Mbps to 100 Gbps. A single, common Application Programming Interface (API) allows application software to be developed once and used with a broad range of Napatech accelerators. This allows combinations of different accelerators with different port speeds to be installed in the same server. Additional features include:

- Napatech accelerators can identify, filter and distribute flows to up to 32 CPU cores
- Data merging functionality allows flows from different ports on different accelerators to be merged for analysis
- Data sharing functionality allows multiple applications to access the same data at the same time
- All of this can be performed with very low server CPU load

CHOOSE THE MARKET LEADER

Napatech is the market leading provider of accelerators for network management and security applications. Napatech provides global sales and support from local offices in all major continents, which is included in the price of the accelerator. This means that our highly experienced support resources are available for design and integration support, as well as field support without extra charge.

Napatech accelerators are manufactured to the highest standards by outsourced manufacturers in Switzerland and the USA supporting all major certifications including NEBS for telecom applications.

COMPANY PROFILE

Napatech is the world leader in accelerating network management and security applications. As data volume and complexity grow, the performance of these applications needs to stay ahead of the speed of networks in order to do their jobs. We make this possible, for even the most demanding financial, telecom, corporate and government networks.

Now and in the future, we enable our customers' applications to run faster than the networks they need to manage and protect.

Napatech. FASTER THAN THE FUTURE
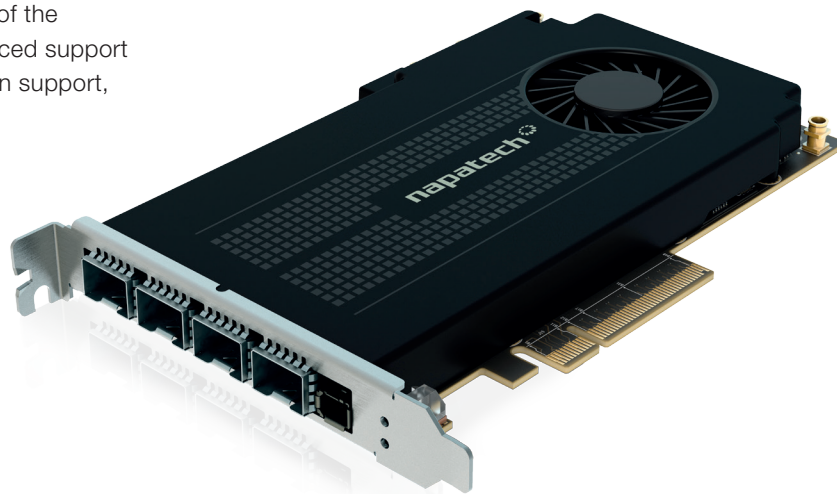
EUROPE, MIDDLE EAST AND AFRICA
Napatech A/S
Copenhagen, Denmark

Tel. +45 4596 1500
ntemeasales@napatech.com
www.napatech.com

NORTH AMERICA
Napatech Inc.
Boston, Massachusetts
Mountain View, California
Washington D.C.

Tel. +1 888 318 8288
ntamericassales@napatech.com
www.napatech.com

SOUTH AMERICA
Napatech
São Paulo, Brazil

Tel. +55 11 2127 0782
ntsouthamericasales@napatech.com
www.napatech.com

APAC
Napatech Japan K.K.
Tokyo, Japan
Tel. +81 3 5326 3374

Napatech Korea
Seoul, South Korea
Tel. +82 2 6001 3545

ntapacsales@napatech.com
www.napatech.com