Evaluating the Impact of Congestion Reduction on Power Consumption in IoT Networks

J.R. Emiliano Leite, Edson L. Ursini, Paulo S. Martins

Abstract— **Congestion reduction is a critical element in almost any system, and especially in those that are battery supplied and provide critical services, such as in an uninterruptible power supply of an emergency room in a hospital. In this work, we analyze the impact of congestion reduction on the network power consumption of battery-dependent devices. We built a discrete-event simulation model of an IoT network, which was analytically validated by Jackson networks, and quantitatively showed through two case studies how power consumption may be decreased in a network node by reducing congestion of the network traffic. The model and its implementation serve as the basis for the analysis of several scenarios that may support the design, planning and dimensioning of future IoT networks regarding traffic congestion. For existing networks, it may also be used for detection or even prediction of future bottlenecks. The analysis of the results indicates that a reduction in power consumption in the overall network is achievable - which may be used to extend the system's lifetime.**

Keywords— **Jackson networks, Discrete event simulation, Ad-Hoc networks, RFID**

I. INTRODUCTION

Traffic congestion plays a critical role in the design of battery-dependent devices and networks, since the networks experience retransmissions at a large rate that may increase power consumption and thus deplete battery sources more quickly. Within this context, it is possible to elicit important design issues for an IoT network: 1) How much traffic congestion can the network tolerate, given a certain QoS requirement? For such level of congestion, we may ask: 2) What is the power consumption in an arbitrary network node (which allows us to establish the duration of its battery/lifetime)? 3) What are the congestion bottlenecks? 4) What is the traffic arrival rate that eliminates such bottlenecks? and 5) What is the processor service rate that eliminates such bottlenecks?

The methodology that seeks to answer these questions - and that is pursued in this work - is the following:

1) Construction of the discrete event simulation (DES) model for the complete system (Fig. 1, Table I, Section III);
2) Validation of the DES model, e.g. using the analytical model (Jackson network) (Section V, [1]);
3) Congested network evaluation: The identification of bottlenecks is a plus in this step (Case 1, Section IV-A);
4) Uncongested network evaluation (Case 2, Section IV-B);
5) Calculation of the variation in traffic (Δ Erlang, Section IV);
6) Estimation of reduction in power consumption from Δ Erlang using the equations/models (Watt/Erlang) for several different technologies (Section V).

The contribution of this work lies in offering an analytically validated discrete-event IoT network simulation model, which includes an AdHoc (Mobile AdHoc Network, MANET) and a RFID network combined. The model can be used to reason about energy saving as a function of traffic congestion. Furthermore, the study also provides case studies that showcase the approach. To the best of our knowledge (and as discussed in Section II), we have not found in the literature review work that approaches this topic with such features.

The remainder of this paper is organized as follows: In Section II, we review previous work. The system model is discussed in Section III. Two case studies illustrating the application of the model are shown in Section IV. In Section V, the results are discussed. We summarize and present our conclusions and future work in Section VI.

II. RELATED WORK

Mobile communications consume a significant amount of energy. In the work by Dahal et. al [2], more than 50% of the total energy is consumed by the radio access, and within this fraction - 50-80% is used for the power amplifier. The results revealed a linear relationship between the power consumption and traffic loads, and the authors provided suggestions for energy-efficient wireless communication.Measurements show the existence of a direct relationship between base station traffic load and power consumption. The paper by Lorincz et al. [3] developed a precise linear power consumption (Watt X Erlang) model for base stations of GSM (Global System for Mobile Communications) and UMTS (Universal Mobile Telecommunications System).

In the work by Deruyck et al. [4], a power consumption model based on network traffic for base stations is proposed. This model is validated by temporal power measurements on actual base stations. The energy efficiency of three different wireless technologies (WiMAX, LTE, and HSPA) is compared.

Ghandi et al. [5] measured the variation of power in relation to the corresponding variation of traffic (in Erlang) in a CDMA network. Other parameters such as the number of lost calls, the number access failures and the duration in minutes were also considered. They found the variation of power to be larger than the variation in traffic. Based on the measured data, they estimated an analytical/empirical model. From the viewpoint of the relation power per Erlang, they found that a network with small cells to be the most effective.

The paper by Hinton et al. [6] presents a network-based model of power consumption for the Internet infrastructure. The access network dominates the Internet's power consumption. However, as the access speed grows, the power

consumption in the core network routers prevails over the access network consumption. Several strategies were created to improve the energy efficiency of the Internet.

Our work differs from previous research in that our model is an IoT network, unlike e.g. Deruyck et al. who deal with both macro-cell and micro-cell base stations, or Hinton et al. who address power consumption in the Internet. Previous work have also taken direct measurements on real physical networks, whereas our work has employed discrete event simulation. Much like previous research, our work also focuses on the relationship between network traffic congestion and power consumption. However, our focus lies in establishing the steps to determine this relationship using discrete event simulation (and analytical models for the purpose of validation). It analyses both congested and uncongested network scenarios in order to establish the level of power reduction. As such, our method has the capability of being used as a management tool for existing networks or as planning and dimensioning tool for future designs.

III. SYSTEM MODEL

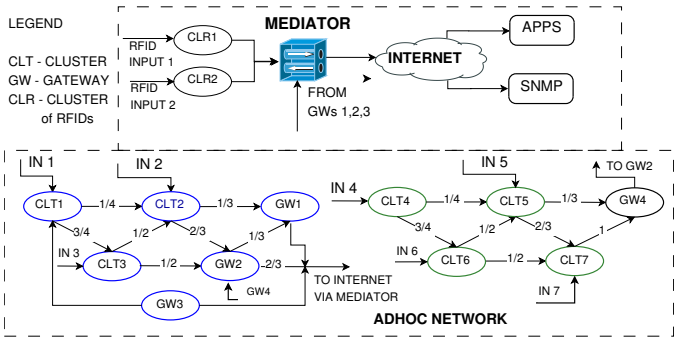

Fig. 1. IoT Network model.

An IP packet is modeled as an entity that arrives to the system and crosses several internal queues in a cluster before its departure (i.e. before it is consumed by an application). The network model is a hierarchy consisting of clusters which contain nodes, which in turn have multiple CPUs, thus allowing several parallel connections. Inherent to each queue is the waiting delay before a packet can be processed by a server. Clearly, both queueing and processing times are subject to statistical distributions. Therefore, a network cluster may be regarded as a set of internal queues (each one associated with an outbound link). The network components are the mediator, gateway and endpoints. The endpoints can be RFID and sensors for different applications. Figure 1 shows the network model with its inputs (packets) and outputs (packets) for each cluster. The upper part of the model consists of the RFID network, the Mediator, Internet and applications. Their details are as follows:

- Mediator *MD*, which is a node and contains one or more CPUs;
- Two clusters (CLR_1, CLR_2), which perform the acquisition/input of RFID tags;
- Two RFID inputs, which receive data packets generated by IoT RFID tags (RFID reader);
- Two applications: 1) smart and green building including the control of actuators (light, auxiliary power supply (UPS), air conditioning), and 2) SNMP management. Note that the path from the application to the actuators, and the actuators themselves are not the focus of this work;
- Internet, which models the traditional Internet.

The lower part of Fig. 1 is an AdHoc Network that generates data traffic which is aggregated by the IoT mediator. It consists of the following elements:

- 7 clusters (CLT_1....CLT_7); these are non-mobile and homogeneous for the sake of simplicity. However, the model does not restrict the addition of heterogeneous clusters. Each cluster consists of up to 2 cluster heads and n mobile nodes/processors, where n is a configurable parameter. Each cluster head is a static (non-mobile) intermediate system, i.e. all the traffic that leaves the cluster is sent through the cluster head;
- 4 gateways or Internet nodes (GW_1....GW_4); Both GW_1 and GW_2 are output gateways; GW_3 is an emergency gateway, i.e. it is used as a backup gateway for GW_1, e.g. when the latter overflows its internal buffers; GW_4 and also GW_2 are protocol converters, i.e. they are used to integrate two subnets;
- 7 Inputs: model data packets generated by IoT sensors;
- 3 Internet outputs (via Mediator): from GW_1, GW_2 and GW_3, they model the flow of IP packets outbound;
- Input variables: data arrival and service time distributions in a node;
- Control variables: probability of node connectivity in a cluster. This probability is provided by the Random Waypoint algorithm (which depends on a range of variables such as receiver threshold, area size, antenna type, height, and gain, and system loss coefficient among others);
- Output variables: mean queue time and mean CPU utilization on each cluster for a given position of the nodes within the cluster.

Each cluster contains several nodes which in turn have internally one or more CPUs. In addition, the output CPUs in each cluster are used for its output channels/links (Table I). These output CPUs (clusters heads) are fixed in our model (without sacrificing the quality of the results), although it is possible to configure them to have some limited degree of mobility as well. Nodes share the output CPUs for relaying outbound traffic, provided that they have connectivity, i.e. they are within the power range of either an output CPUs or an intermediate node.

Each node is modeled as four simulation blocks connected in series: *Enter* block: the enter block simulates the arrival of a packet in a cluster. It counts the number of packets entering the cluster; *Decide* block distributes the packets across a set of outgoing lines, where each line is associated with an outgoing queue. An important parameter in this block is the probability of packet loss (due to connectivity loss as given by the Random Waypoint, RWP, model), and its value was obtained from the case study (Section IV). The probabilities of a packet being forwarded to an outgoing link are initially configured

TABLE I
NETWORK CONFIGURATION.

Function	Probability	Output CPUs
GA_{12}	1	25
GA_{34}	1	28
GA_{57}	1	29,33
GA_{8}	1	19,34
AdHoc submodel	1	30
AdHoc submodel	1	31
AdHoc submodel	1	32
CLR_1	1	21
CLR_2	1	22
MD	1/2, 1/2	24, discard
GWI	1/3,1/3,1/3	23, 26, 27
GW_1	1	5
GW_2	1/3, 2/3	11,6
GW_3	1	14
GW_4	1	17
CLT_1	1/4, 3/4	1,2,20*
CLT_2	1/3, 2/3	3,4
CLT_3	1/2, 1/2	7,8
CLT_4	1/4, 3/4	12,15
CLT_5	1/3, 2/3	16,13
CLT_6	1/2, 1/2	9,10
CLT_7	1	18

* output-CPU 20 is used only in an emergency
GA - application gateway

as shown in Fig. 1 (e.g. 1/4 from cluster 1 to cluster 2 and 3/4 from cluster 1 to cluster 3); *Output queue* represents the queueing time in the outgoing line; *Output cluster* simulates the output (i.e. forwarding) of packets from the cluster. It is also responsible for counting the number of packets leaving the cluster.

Table I shows the relation of cluster / gateways to output CPUs. The column "Probability" is associated to the column "Output CPUs". Each probability is used to define the traffic management of each node according to a given application. These values also indicate the probability of a packet being serviced by the indicated output CPU. For example, the probability that cluster CLT_2 sends a packet to output CPU_3 is 1/3, and this probability is 2/3 for output CPU_4.

Each node receives packets at the input link and forwards them to one of the outbound links using UDP over IP (Datagram). Since the arrival of requests for the RFID and AdHoc networks can be modeled as a Poisson process, the traffic volume of each individual node can be extended to the traffic volume of a cluster by the simple sum of the rates of Poissonian arrivals. Thus, we sum the rates of each node to form a cluster of ten nodes, for instance.

The model adds two representative applications that process and consume the information leaving the mediator, SNMP management, and Smart Green. Smart green applications receive and store information in databases. RFID information is all based on traffic generated by RFID tags (96-bit EPC-Global). The SNMP application receives a trap from the mediator, and it sends a request to the same, which replies with a response. In typical IoT, the databases can be reached by mobile or cell phone applications through the secure HTTPS protocol.

IV. CASE STUDY

The results for the two case studies are presented in Table IV and discussed in the following subsections.

A. Case 1 - Congested Network

Fig. 2 illustrates the results for Case 1. Case 1 (Table IV) shows the case for a congested network. It uses an arrival rate of EXPO (0.4) (= 2.5 packets/sec). We also need the 14 x 14 matrix R, which describes the probabilities shown in Fig. 1, since it is used to evaluate the degradation of the nodes that are congested.

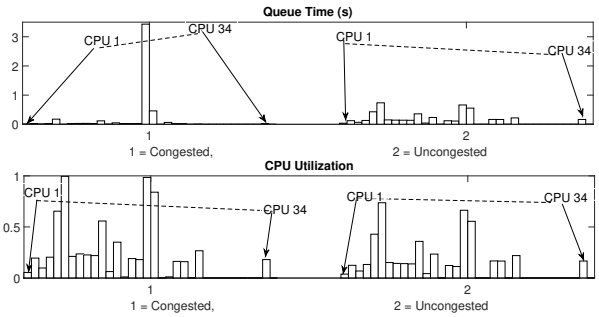

Fig. 2. CPU mean queuing time and utilization: Cases 1 and 2.

Table IV shows the total average message delay times (in seconds). The delays depend on the path taken by the packets in the routing scheme. The first column shows the type of messages, which are 1) SNMP messages; 2) Sensor 1 measures current and voltage for an air conditioning (A/C); 3) Sensor 2 indicates level of illumination; 4) Sensor 3 reads current and voltage battery values (UPS), and 5) RFID tag.

The message delays are measured from their source inputs (clusters) to the MD outputs. Tables II and III are related to this case, i.e. the evaluation of traffic under congestion. The values of γ refer to the packet generation rate in each cluster, according to Table II. From the values of γ and Matrix R, we may calculate the λ values that correspond to a Jackson network [7]. Further details of this calculation are shown in the following subsection.

In Fig. 2 we observe that some CPUs are overloaded (i.e. high CPU utilization), and the overloading of a single node may be propagated to other nodes to the point of compromising the whole network (in the event of a worst case scenario). Due to the fact that the mean queue time of CPU 6 is very high (39 s), it was not placed in Fig. 2 (congested, top/left) in order to allow the other CPUs to be viewed.

This congestion leads to higher CPU utilization in each node since congestion triggers the error detection and correction as well as retransmission of lost packets. It is only in the following case, i.e. Case 2, that we decongest the network and the difference in the utilization is related to the reduction in power consumption for the batteries of the mobile devices. By figuring out this relationship, we become able to estimate the actual gain in power consumption. Considering that the capacity of both CPUs (except MD CPU) and links between clusters is constant, we reduced the arrival rates to decongest

TABLE II

CONGESTED (CASE 1) AND UNCONGESTED (CASE 2) NETWORK TRAFFIC - ARRIVAL RATES (γ AND λ) IN PACKETS/SEC

Case 1														
—	CLT1	CLT2	CLT3	CLT4	CLT5	CLT6	CLT7	GW1	GW2	GW3	GW4	CLR1	CLR2	MD
γ	2.5	2.5	2.5	2.5	2.5	2.5	2.5	0	0	0	0	1.67	1.67	5.0
λ	2.5	5.31	4.38	2.5	5.31	4.38	8.23	7.01	15.7	0	10.0	1.67	1.67	25.8
Case 2														
—	CLT1	CLT2	CLT3	CLT4	CLT5	CLT6	CLT7	GW1	GW2	GW3	GW4	CLR1	CLR2	MD
γ	1.67	1.67	1.67	1.67	1.67	1.67	1.67	0	0	0	0	1.67	1.67	5.0
λ	1.67	3.54	2.92	1.67	3.54	2.92	5.49	4.68	10.49	0	6.67	1.67	1.67	20

TABLE III

CONGESTED NETWORK (CASE 1) AND UNCONGESTED (CASE 2) TRAFFIC (ERLANG) PER CPU

Case 1 (Total = 6.734 Erl)												
CPU	1	2	3	4	5	6	7	8	9	10	11	12
Traffic (Erl)	0.125	0.125	0.267	0.267	0.71	0.587	0.219	0.219	0.219	0.219	0.587	0.125
CPU	13	14	15	16	17	18	19	20	21	22	23	24
Traffic (Erl)	0.267	0	—	0.267	1.00	0.823	–	–	0.167	0.167	–	0.258
Case 2 (Total = 4.896 Erl)												
CPU ID	1	2	3	4	5	6	7	8	9	10	11	12
Traffic	0.084	0.084	0.177	0.177	0.468	0.525	0.146	0.146	0.146	0.146	0.525	0.084
CPU ID	13	14	15	16	17	18	19	20	21	22	23	24
Traffic	0.177	0	0	0.177	0.667	0.549	0	–	0.167	0.167	–	0.20

CPU 19 - RESERVED (for future extensions), CPU 20: Emergency, CPU 23, 25 to 34: Application

the network, as shown in the next subsection. To verify that the network was decongested, we used Jackson networks.

B. Case 2 - Uncongested Network

Case 2 (Table IV) shows the simulation results where the arrival rate is set to an EXPO (0.6) (= 1.67 packets/s) for a stable (non-congested) network. Three types of messages were considered, i.e. SNMP control, RFID, and sensor messages (light, AC and battery). Tables II and III are related to Case 2 and show the evaluation of traffic under uncongested load. Figs 2 shows the results for Case 2.

Since the initial simulation model has both exponential arrival and service distributions, it may be validated against Jackson's open queuing network model [7]. The solution is obtained from a Markov chain. The packet arrival rate is 1/0.6 = 1.67 packets/s. The first seven arrivals, each generated by a cluster (gateways do not generate traffic), yield 1.67 packets/s (the remaining four are gateway inputs), therefore: γ = [1.67, 1.67, 1.67, 1.67, 1.67, 1.67, 1.67, 0, 0, 0, 0, 1.67, 1.67,5.0]. We also need the 14 x 14 matrix R, which describes the probabilities shown in Fig. 1 . The total arrival rates in each cluster or gateway is given by the vector: $\lambda = \gamma\ [I - R]^{-1}$, λ = [1.67, 3.54, 2.92, 1.67, 3.54, 2.92, 5.49, 4.68, 10.49, 0, 6.67, 1.67,1.67,20]. From the rates obtained from Table II, it is possible to calculate the packet delay for each CPU (W_i, [i=1....24]) by means of the equation $W_i = \frac{\lambda_i/\mu_i}{\mu_i - \lambda_i}$, which gives the delay in an M/M/1 queue, where λ_i and μ_i are the rates for each CPU, and $\mu_i = \frac{1}{0.1} = 10$ packets/s. It is also possible to calculate the traffic, in Erlangs, for each CPU as $\frac{\lambda_i}{\mu_i}$ (Table III). Since all the delay values obtained from the simulation model matched the ones from the analytical model, the simulation model may be deemed validated. This validation is a crucial step since it allows further extensions to this model, i.e. the inclusion of other model features such as new types of distributions. Due to the high utilization of the mediator (output CPU 24), its service rate was increased 5 times (i.e. from 10 to 50 packets/s).

The initial distribution adopted for the arrival and service rate was the exponential. This distribution is suitable since 1) it allows the validation of the model with an analytical model; 2) it is the one that stresses the network (the worst-case when there is no bursts). If the exponential distribution does not match the reality, it is possible to combine exponential distributions to form Erlang(k) distributions, which may better reflect and the actual traffic model in the network. Otherwise, if there are bursts in network, the Pareto or Hyper-exponential distributions may be employed, depending upon the application. Once the model is validated by incremental evolution, other types of extensions and distributions may be studied.

TABLE IV

MESSAGE DELAY TIME (SECS) FOR CASE STUDIES.

Message type	Case 1 congested	Case 2 uncongested
SNMP	0.093	0.085
sensor 1 AC sensor 2 light sensor 3 battery	28.04	2.28
RFID	2.170	2.145

V. ANALYSIS, REMARKS AND DISCUSSION

It was possible to realize through this work that the methodology allowed us to identify the bottleneck and by increasing its capacity, it was possible to reduce congestion in the network.

The impact of congestion reduction on power consumption depends on the hardware implementation. Table V shows scenarios taken from six typical networks. For example, in the work by Dahal et al. [2], power consumption for base stations in ten consecutive days (including weekends), and for 864000 samples collected from a 3G system - is given by $y = a + bx$, where a is given in Watt and b in Watt/Erl; the value of x is the reduced traffic in Erlangs. Under high traffic, y = 1.274 + 1.713x, and the regression has a coefficient of determination reasonable but not large, i.e. $r^2 > 0.765$ (Table V, 3G - BS).

TABLE V

IMPACT OF TRAFFIC REDUCTION ON POWER CONSUMPTION.

Equipment [ref]	Equations (W)	Busy Hour Reduction (Watts)	
		Jackson	Simul.
3G - BS [2]	1.274 + 1.713x	4.42	4.94
GSM900 - sector1 [3]	581 + 11.9x	602.9	606.5
GSM900 - sector2 [3]	549 + 11.1x	569.4	572.8
UMTS - 3 sectors [3]	551 + 1.14x	553.1	553.4
WiMAX, LTE, HSPA [4]	493.2x	906.5	1055.4
CDMA (forwarding link) [5]	0.734x	1.35	1.57

If our simulation model had a 3G implementation such as the one by Dahal et al. [2], and considering that the values we obtained are for high (or peak) traffic, we would estimate a daily reduction in power consumption of 4.42 Watts for 24 CPUs (i.e. 1.274 + 1.713 (6.734-4.896) = 4.422 W). For the simulation model, considering all 34 CPUs, the energy saving is 1.274 + 1.713 (7.056-4.915) = 4.94 W. The difference of 10 CPUs, besides statistical fluctuations in the analytical and simulation model may explain the small difference (i.e. 4.42 and 4.94). It is an indication of validation of the models, although a more strict validation procedure was developed by Leite et al. [1]. For the highest traffic load between 8-11 am to 6-8 pm, these figures would roughly translate into a monthly savings of up to 450 Watts · h/month in the whole network, which could - in turn - represent a substantial extension of battery life. The calculation of the impact for the other networks follows the same reasoning. This implies in about 100 times (5 hours per day times 20 days per month) the values of columns 3 and 4 in Table V.

To simulate the performance of the network, the adopted mobility model was the Random Waypoint (RWP). To evaluate each node independently, a MATLAB routine generates random positions for the ten nodes within each cluster every one second. Lastly, the proposed model is general and it can be instantiated for specific applications. For example, the probabilities of transmission for outgoing links can be measured in a real application and replaced in the model. The arrival and service distributions considered may also be replaced by actual measurements and/or other types distributions.

VI. SUMMARY AND CONCLUSION

In this work, we addressed the impact of reduction of network congestion on power consumption. This was carried out by following a methodology involving six steps as outlined earlier. We tackled traffic congestion since it is a critical component that adds power consumption and thus reduce battery lifetime of both devices and communication infrastructures.

To illustrate the approach, we built a discrete event simulation model that included an AdHoc network, a mediator, a set of applications and a set of inputs (RFID, sensors) which generate traffic input to the network. The simulation model was then exercised through two case studies that estimated the network traffic both with and without congestion. The network traffic then may be translated into power consumption via an energy consumption model that depends on the target implementation hardware platform.

The model is capable of estimating the power consumption both globally and at individual nodes, and both at design and operation time. This methodology is, at design time, the only viable alternative means to estimate power consumption on large networks, since measurements and data from the physical network are not available, and since analytical models cannot capture the complexity of such networks. For an already existing and operational network, we argue that it is also the easiest approach to estimate future bottlenecks and predicting power consumption due to network expansion (growth).

As future work, we consider that the reduction of power consumption may be better explored by adding a central entity for traffic management implemented as a fuzzy logic controller.

REFERENCES

[1] J. R. E. Leite, E. L. Ursini, and P. S. Martins, "A proposal for performance analysis and dimensioning of iot networks," in *Brazilian Technology Symposium (BTSym 2017)*, December 2017.

[2] M. S. Dahal, S. K. Khadka, J. N. Shrestha, and S. R. Shakya, "A regression analysis for base station power consumption under real traffic loads - a case of nepal," *American Journal of Engineering Research (AJER)*, vol. 4, no. 12, pp. 85–90, 2015.

[3] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modelling of base station power consumption under real traffic loads," *Sensors*, vol. 12, no. 4, pp. 4281–4310, 2012. [Online]. Available: http://www.mdpi.com/1424-8220/12/4/4281

[4] M. Deruyck, W. Joseph, and L. Martens, "Power consumption model for macrocell and microcell base stations," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 3, pp. 320–333, Mar. 2014. [Online]. Available: http://dx.doi.org/10.1002/ett.2565

[5] A. D. Gandhi and M. E. Newbury, "Evaluation of the energy efficiency metrics for wireless networks," *Bell Labs Technical Journal*, vol. 16, no. 1, pp. 207–215, June 2011.

[6] K. Hinton, J. Baliga, M. Feng, R. Ayre, and R. S. Tucker, "Power consumption and energy efficiency in the internet," *IEEE Network*, vol. 25, no. 2, pp. 6–12, March 2011.

[7] J. R. Jackson, "Networks of waiting lines," *Operations Research*, vol. 5, no. 4, pp. 518–521, 1957.