Vehicular Communications 9 (2017) 162–175

Contents lists available at ScienceDirect

Vehicular Communications

www.elsevier.com/locate/vehcom

Energy efficient Nano Servers provisioning for Information Piece Delivery in a vehicular environment

Samaneh Igder, Samya Bhattacharya *, Bilal Qazi, Hamdi Idjmayyel, Jaafar M.H. Elmirghani

School of Electronic and Electrical Engineering, University of Leeds, UK

ARTICLE INFO

Article history:
Received 30 November 2016
Received in revised form 18 May 2017
Accepted 18 June 2017
Available online 23 June 2017

Keywords:
Information Piece Delivery
Mixed integer linear programming
Sleep cycles
Rate adaptation
Nano Servers (NS)

ABSTRACT

In this paper, we propose energy efficient Information Piece Delivery (IPD) through Nano Servers (NSs) in a vehicular network. Information pieces may contain any data that needs to be communicated to a vehicle. The available power (renewable or non-renewable) for a NS is variable. As a result, the service rate of a NS varies linearly with the available energy within a given range. Our proposed system therefore exhibits energy aware rate adaptation (RA), which uses variable transmission energy. We have also developed another transmission energy saving method for comparison, where sleep cycles (SC) are employed. Both methods are compared against an acceptable download time. To reduce the operational energy, we first optimise the locations of the NSs by developing a mixed integer linear programming (MILP) model, which takes into account the hourly variation of the traffic. The model is validated through a Genetic Algorithm (GA1). Furthermore, to reduce the gross delay over the entire vehicular network, the available renewable energy (wind farm) is optimally allocated to each NS according to piece demand. This, in turn, also reduces the network carbon footprint. A Genetic Algorithm (GA2) is also developed to validate the MILP results associated with this system. Through transmission energy savings, RA and SC further reduce the NSs energy consumption by 19% and 18% respectively, however at the expense of higher download time. MILP model 4 (with RA) and model 5 (with SC) reduced the delay by 81% and 83% respectively, while minimising the carbon footprint by 96% and 98% respectively, compared to the initial MILP model.

1. Introduction

In a city vehicular environment, mobile users in the vehicles are one of the main driving forces behind traffic growth. In the recent past, researchers predicted an 8 fold growth in traffic by 2020 [1]. Energy efficiency in the vehicular domain (both vehicle to vehicle and vehicle to infrastructure) has been widely researched. While the bulk of the savings is achieved through optimising the locations of fixed nodes like base stations (BSs) or roadside units (RSUs), further savings may be possible by reducing the transmission energy consumption of a node by introducing sleep strategies during the inactivity periods [2–4]. These techniques have already been utilised for the line-cards in the routers [3–5], where up to 79% reduction in energy consumption was achieved. Introducing sleep cycles is an attractive solution for vehicular networks, as it does not require a complete overhaul of network devices, protocols or architecture. However, major savings (as in the case of wired networks) may not be feasible in vehicular networks as they are not intrinsically over-provisioned and the link quality depends upon the varying wireless channel, which makes such systems susceptible to degraded Quality of Service (QoS). Nevertheless, a few research groups have proposed a number of sleep strategies to make cellular networks energy efficient [6,7]. In a macro–micro cellular architecture [8], where small RSUs are used for offloading purposes, introducing sleep cycles can be extremely effective due to the shorter resource activation time of an RSU. Therefore, such mechanisms are worth exploring in the city vehicular information piece delivery scenario.

The authors in [8] proposed a sleep strategy which reduced the energy consumption of the RSU by switching OFF only its transmitting circuitry part for a randomly distributed time duration when there is no request to serve. The RSU remained in sleep mode for a fixed time duration (randomly generated with a certain mean value) even if pieces were waiting to be served. Upon waking up, the RSU served the pieces that arrived (if any) and switches

* Corresponding author.
E-mail addresses: elsi@leeds.ac.uk (S. Igder), s.bhattacharya@leeds.ac.uk (S. Bhattacharya), bilalqazipk@gmail.com (B. Qazi), h_jmayyel@hotmail.com (H. Idjmayyel), j.m.h.elmirghani@leeds.ac.uk (J.M.H. Elmirghani).

http://dx.doi.org/10.1016/j.vehcom.2017.06.001

back to sleep mode when the buffer becomes empty. This mechanism was called Random Sleep Cycles. This type of sleep cycles degrades the system performance, when a large number of information pieces wait to be served. Since energy saving through sleep cycles was achieved at the expense of degraded QoS and incurred wake-up overheads, associated with each sleep cycle [8], there is a need to improve QoS and reduce wake-up overhead while maximising energy savings. One way of achieving this is to put a limit (bound) on delay which helps determine the maximum sleep duration for a given period (we consider one hour periods). Maximising the average sleep duration for each hour in turn minimises the wake-up overhead. Another technique, called 'Rate Adaptation', has traditionally been used by the network operators to dynamically allocate bandwidth to the end users based upon the quality of the wireless channel between the BSs and end users [9,10]. Hence, the user with good channel is allocated higher bandwidth compared to the user with poor channel conditions. However, to the best of our knowledge, rate adaptation has never been used in the context of saving non-renewable energy, or improving QoS, if renewable energy is available. In [11], traffic sensing based data load adaptation was proposed in a city vehicular scenario, which inspired us to formulate the concept of energy aware rate adaptation, proposed in this paper. The load adaptation is based on traffic sensing in real time, which guides us to vary the data rate of a device. In this paper, our proposed rate adaptation technique adjusts the service rate of the Nano Servers (NSs) according to the available energy. The service rate varies linearly with the available energy within a range of 3 Mbps (Minimum) to 27 Mbps (Maximum). Note that the energy used by the transmitter may not scale up linearly with the data rate, especially if multiple antennas are used [12]. Rate adaptation under such conditions would be considered in future work.

The main challenge of using renewable energy in Information Piece Delivery (IPD) lies in its judicial distribution in order to achieve best QoS. Invariably, this requires detailed knowledge of the traffic patterns. To the best of our knowledge, transient performance analysis of NSs using sleep cycles, rate adaptation, and renewable energy in vehicular IPD has not been done before, and therefore forms the main contribution of the paper. The energy aware rate adaptive NSs introduced in this paper operate at variable service rate according to the available transmission energy, while sleep enabled NSs switch off the transmission (networking) circuitry part to save energy. We developed five models: Model 1, Model 2, Model 3, Model 4 and Model 5 that form the contributions of this paper:

1. Model 1 minimises the energy consumption of IPD by optimising the number and location of the Nano Servers (NSs) while accounting for vehicle mobility.
2. Model 2 and Model 3 reduce the non-renewable transmission energy consumption of the IPD by introducing random sleep cycles and energy aware rate adaptation, respectively while maintaining the minimum acceptable QoS.
3. Model 4 improves the overall network QoS and reduces the carbon footprint by optimally distributing the available renewable energy according to the piece demand at each NS, hence reducing the waiting delay.
4. Model 5 is an extension of Model 3. It improves the overall network QoS and reduces the carbon footprint by optimally distributing the available renewable energy according to the information piece demand at each NS. It reduces the service delay, which in turn also reduces the waiting delay.

The rest of the paper is organised as follows. In Section 2, we describe a city vehicular environment for information piece delivery. Section 3 presents the formulation of Model 1 for optimising the locations of ordinary (non-rate adaptive) NSs thereby reducing the overall network energy consumption. Section 4 describes the formulation of a service rate optimisation constraint using a queueing model for the proposed rate adaptive NS. The corresponding model (Model 2) minimises the Non Renewable Energy (NRE) consumption while maintaining an acceptable service rate. Section 5 describes Model 3, which optimises the distribution of Renewable Energy (RE) for rate adaptive NSs to further reduce NRE while achieving the best possible QoS. Section 6 presents the results with discussion. Finally, the paper is concluded in Section 7.

2. Related work

In a smart city vehicular environment, mobile users in the vehicles are main driving force towards the growth in traffic. Mobile traffic is projected to grow with a compound annual growth rate of 47% between 2016 and 2020 [9]. Researchers expected eight times growth of traffic by 2020 [1]. Hussain et al. [13] coined the idea of vehicular clouds by taking traditional VANET to the cloud. Vehicular cloud computing is a new hybrid technology solution with impact on the Intelligent Transportation System (ITS). It uses the resources of vehicles such as GPS, storage, processing, and Internet connectivity as well as information sharing with the cloud [14]. The European project TROPIC [15] pursued the idea of bringing cloud capabilities closer to the mobile devices. A cloud brought close to the ground to support smart things is referred to as Fog computing, which is a concept introduced by Cisco [16]. The service can be provided by the Nano-Servers (NSs), which are usually positioned between the cloud and smart devices. These can be used for content delivery as well as other information-based services in the city.

Vehicular Cloud Computing (VCC) is emerging as a new paradigm with the goal of merging mobile cloud computing with high-speed vehicles, where the vehicles have on board devices. The primary focus is to include safety applications, as well as the emerging future Internet applications [5]. These may be content delivery or infotainment applications (such as, Netflix and YouTube). In [17], the authors optimise the throughput of the secondary users of vehicular networks, while respecting the quality of service (QoS) requirements of the primary users. Throughput is evaluated while taking into account the maximum tolerated collision rate with the secondary users. This gave rise to a new optimised medium access (MAC) protocol, which is able to maximise the average aggregated throughput of the system. Avatar [18] is a system that utilises cloud resource to support fast, scalable, reliable, and energy efficient distributed computing over mobile devices. It is a virtual machine per-user in the cloud that runs applications on behalf of the user's mobile devices. The system is essentially distributed and synchronised, which requires: creating a high-level programming model and a middleware that enables effective execution of distributed applications.

Energy efficiency in the vehicular domain (both vehicle to vehicle and vehicle to infrastructure) has been addressed through optimising the location of fixed nodes such as base stations (BSs) and roadside units (RSUs) to reduce network energy consumption [19–21]. Further savings may be possible by reducing transmission energy consumption at a node by introducing sleep strategies during the inactivity periods [2,3]. Introducing sleep cycles is an attractive solution for wireless networks, as it does not require a complete overhaul of network devices, protocols or architecture. Such techniques have already been studied for the line-cards in the routers [4,6] where up to 79% reduction in energy consumption was achieved. Such major reduction may not be feasible in wireless and mobile network (e.g. cellular or vehicular) as they are not intrinsically over-provisioned and the link quality is dependent upon the varying wireless channel, which makes it susceptible to degraded quality of service. Nevertheless, a few research groups

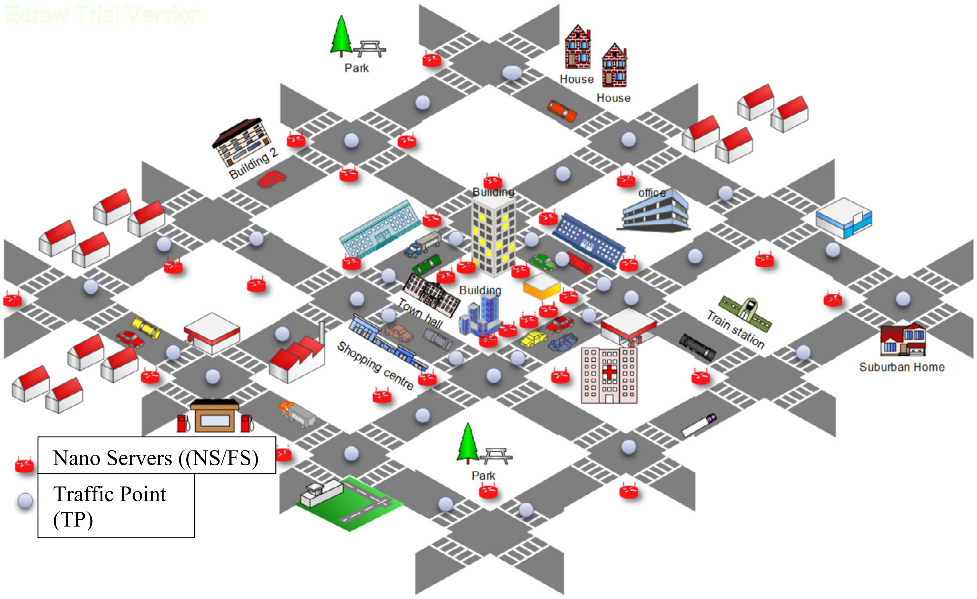

Fig. 1. Smart city vehicular environment.

have proposed a number of sleep strategies to improve cellular networks energy efficiency [10,22]. In [23], the authors proposed dynamic switching for a BS in low traffic conditions. However, fast switching may not be feasible to accommodate transient traffic behaviour because of the number of operations a large BS has to perform [24]. In a macro–micro cellular architecture [25], where small RSUs are used for offloading purposes, introducing various types of sleep cycles (random [26] and adaptive [27]) can be extremely effective due to the shorter resource activation time of an RSU. Therefore, such mechanisms are worth exploring in the context of a smart city vehicular environment.

The authors in [28] proposed a sleep strategy which reduced energy consumption of the RSU by switching OFF only its transmitter part for a randomly distributed time duration when there is no request to serve. The RSU remained in sleep mode for a fixed time duration (randomly generated with a certain mean value) even if contents were waiting to be served. Upon waking up, the RSU served the arrived contents (if any) and switched back to sleep mode when the buffer became empty. This mechanism was called random sleep cycles. This type of sleep cycles degrades the system performance, when a large amount of content is waiting to be served. Since energy saving through sleep cycles type-I was achieved at the expense of degraded QoS [7] and incurred wake-up overhead, associated with each sleep cycle [28], there is a need to improve QoS and reduce wake-up overhead while maximising energy savings. One way of achieving this is to put a limit (bound) on delay, which enables one to find a maximum average sleep duration for that period (i.e. an hour in our case). Maximising the average sleep duration for each hour in turn minimises the wake-up overhead. To improve the performance of vehicular networks, a number of rate adaptation techniques have been investigated in the literature [29–32]. An exhaustive experimental evaluation of rate adaptation algorithms in real environments was presented in [29], followed by the development of a low-overhead rate adaptation algorithm which maximised the network throughput while minimising the bit error rate. The authors in [33] analysed the performance of rate adaptation techniques based on the concept of 'coherence time' using a channel emulator. Moreover, a rate selection policy was presented based on the speed and location of a vehicle [33]. Rate adaptation concepts can also be used for energy efficiency, where the service quality varies according to the available energy. This can be referred to as energy aware rate adaptation. Due to environmental and economic factors, the use of renewable sources of energy such as wind energy is attractive in vehicular networks as well. The main challenge of using renewable energy for content delivery in vehicular networks lies in its judicial distribution for the best achievable QoS. Invariably this requires detailed knowledge of the vehicular traffic patterns and therefore the data traffic patterns they generate. To the best of our knowledge, the time variant performance analysis of NSs using sleep cycles, rate adaptation, and renewable energy in the case of vehicular content delivery has not been studied.

3. Smart city vehicular scenario

The vehicular scenario considered in this paper is one associated with a smart city, where vehicle movements follow a Manhattan Mobility Model [34]. The city area considered is 3 km × 3 km, where there are 24 bi-directional roads and 16 junctions as shown in Fig. 1.

There are a total of 398 possible locations (at roadsides and road junctions) i.e. candidate sites (CS) where the NSs can be installed. The distance between consecutive CSs is 100 m. To cater for the mobility of the vehicles, we represent the set of nearby vehicles as a traffic point (TP) i.e. a centroid. Thus, all the vehicles in the city can be represented by a number of TPs. This means that if a vehicle moves from one TP to another TP, the number of the vehicles in the corresponding TPs changes. In our city scenario the distance between successive TPs is 200 m. If a large number of such TPs is adopted, the impact of discretisation is reduced, however the optimisation computational complexity increases, for example the complexity associated with optimising the number and locations of NSs. Each bi-directional road has five TPs, giving rise to a total of 112 TPs in this scenario. Since the city roads are narrow, the width of the roads is not considered. A cross section of the city is shown in Fig. 2. A vehicle can connect with any of the NSs in range (NS1–NS6), refer to Fig. 2.

The maximum connectivity range is 200 m. We consider short range WiFi communication for information piece delivery for which the NSs can potentially be placed 100 m apart. Furthermore, the communication modules of the vehicles are considered to be perfectly power controlled. This enables overlapping coverage of the NSs. Typical UK city (e.g. Leeds) speed limit is 30 mph (+3 mph tolerance) i.e. 15 m/s. Thus, even if a vehicle travels at

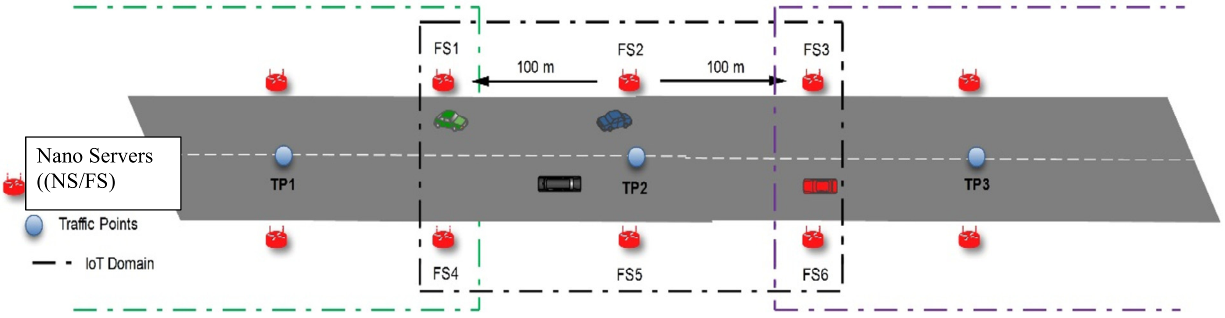

Fig. 2. NSs' coverage range.

this speed limit, the time to go out of range of the neighbouring NSs (dwell time) is 13 s.

Smart things in the city such as commercial premises, healthcare centres, educational buildings, and petrol stations are connected to the NSs in their domain. The NSs are connected to the external network via cloud connections. Each information piece is 2 MB, which contains piggybacked information of the smart city such as environmental monitoring data, flyers, bed and breakfast information, car parking, service station information, medical and healthcare systems information, short movie trailers, advertisements etc. from the city. A vehicle requests a piece from any of the neighbouring NSs in a piecewise fashion to avoid the complexity of handoff management, such that a piece should be downloaded while the vehicle is in range (within the black box in Fig. 2). The NS is assumed to accept a maximum of 9 connections, which implies that each piece is downloaded at 3 Mbps service rate (at least) out of the maximum available 27 Mbps data rate of a NS. Thus, the download time is 3.5 s, which is well within the dwell time (13 s).

The vehicular traffic is obtained from the city of Leeds, UK by averaging and scaling the weekdays, Saturdays and Sundays traffic. The number of vehicles in our city area varies between 16 and 500. Upon reaching a junction, a vehicle chooses one of the four directions with certain probabilities. Out of 4 directions at each junction, two directions lead towards the city centre (in bound) and the other two lead out of the city (out bound). The probability of choosing a direction towards the city centre is variable depending upon the time of the day (0.1–0.9). Considering typical business traffic, we assume the direction probabilities as follows. Between 00:00 hour and 06:00 hour, 50% vehicles are inbound and 50% vehicles are outbound. During 07:00 hour to 09:00 hour, 90% vehicles are inbound (peak hours) and 10% vehicles are outbound. Between 10:00 hour and 17:00 hour, 70% are in-bound and 30% are out-bound. Between 18:00 hour and 21:00 hour, 10% are inbound and 90% are outbound (off peak hours). Finally, from 22:00 hour to 23:00 hour, 50% are inbound and 50% are outbound. Each vehicle generates information piece requests, which follow a Poisson distribution with a mean inter arrival time (τ) of 16 s. The piece size is 2 MB (fixed). The maximum number of cars in an hour in the city is 500. If each car generates a piece request of 2 MB every 16 s, the traffic demand becomes 1 Mbps per car. Hence, the maximum traffic demand should always be less than 500 Mbps. A bespoke JAVA-based city vehicular simulator is developed, which mimics the traffic pattern and also creates the neighbourhood list of TPs for each CS. Each request generated by a vehicle is allocated to the nearest TP and the time stamp is recorded.

3.1. Power consumption profile of a Nano Server (NS)

The maximum power consumption of the proposed rate adaptive NS (Cisco 829 industrial router) is 30 W [35], where 23 W out of the total power consumption is the fixed operational power consumption (consisting of storage and computing services). The computing services consume CPU resource (cycles). Furthermore, 7 W is the maximum transmission power consumption (consisting of networking and some CPU resource). The service rate of the rate adaptive NS linearly varies with the available transmission power [36] between a minimum (3 Mbps) and a maximum (27 Mbps). Without rate adaptation, the NS consumes 30 W and operates at a maximum service rate of 27 Mbps [37]. The NSs are grid connected and therefore have access to both NRE and RE. The analytic wind energy model employed is described as follows. We estimate the average wind power generated by a typical wind farm with 5 turbines in a 24 hour period considering the wind speeds in each hour for a whole year using the formulation of [38]. We assume that 1% of the total wind power generated is available as RE to power the NSs. The parameters for computing wind power are: turbine swept area (30 m^2 [39]), Air density at 15 °C (1.225 $\mathrm{kg/m}^3$) [40], total number of turbines in the wind farm (5), tower height (50–70 m [39]). The renewable wind energy is location based and sporadic (transient), which is disadvantageous and needs to be grid connected for consumer usage. It produces zero carbon footprint. The non-renewable grid energy has continuous supply but produces a carbon footprint. Thus, a proper mix of both energy sources is needed, which will not only maintain the communications quality of service but also reduces the carbon footprint for varying time of the day. This gives rise to a challenging optimisation problem, which we solve through Mixed Integer Linear Programing (MILP).

4. MILP model for transient traffic based Nano Server location optimisation

As mentioned in Section 2, there are a total of 398 possible locations i.e. candidate cites (CS) where NSs can be installed. We develop a mixed integer linear programming (MILP) model, which chooses hourly sets of optimised locations $CSopt_t$ (i.e. $CSopt_t \subseteq CS$) from all possible locations (CS). The MILP model considers vehicle mobility and data traffic transient behaviour through TPs to yield time dependent optimised locations of the NSs (transient optimisation). The union over all time dependent optimised locations ($\bigcup_t CSopt_t$) would give rise to the set of active CS, where NSs need to be installed. This also minimises the overall power consumption of the NSs. Note that the NSs here are of non-rate adaptive type and are energy aware, which means that they operate at full power regardless of the traffic. The formulation sets, parameters and variables for all the MILP models are defined in Table 1.

The traffic demand matrices are used as inputs to the proposed MILP model to find the optimum locations of the NSs. The MILP model is described below.

The proposed model is an information piece distribution scenario. Hence, the vehicles do not generate information pieces and upload to the NSs. NSs receive information piece requests from

Table 1
List of notations and nomenclature.

	Description
Set	
TP	Set of traffic points
CS	Set of candidate sites
NS	Set of installed Nano Servers
$N[j]$	Set of neighbouring TPs for NS n
$N[n]$	Set of neighbouring NSs for TPs j
T	Set of time points within one hour (600 s each)
MILP parameter	
$Bmax$	Maximum capacity of a NS (27 Mbps)
$Ndmax$	Maximum number of simultaneous downloads from a NS (9)
$drmax_{nt}$	Maximum data rate at NS n at time t (i.e. 27 Mbps)
$Dmax_n$	Maximum acceptable average piece delay at NS n (i.e. 13 s)
$Pmax_{nt}$	Maximum power consumption of a NS n at time t (30 W)
$Pidle_{nt}$	Operational power consumption of NS n at time t ($\cong$ 23 W) [41]
$Ptxmax_{nt}$	Maximum transmission power consumption of a NS n at time t given by $Pmax_{nt} - Pidle_{nt} = 7$ W [28]
Pw	Hourly available wind power
Pw_n	Portion of wind power consumed by NS n
$Pmin_{nt}$	Minimum required power by NS n at time t
$Esmaxs_n$	Maximum energy saving of NS n with random sleep cycles
$Psmaxs_n$	Maximum power saving of NS n with random sleep cycles
Ewo	Wake-up overhead energy (in Joules)
λ_{jt}	Traffic demand at TP j at time t
λc	Total data traffic demand at NS n
A	A constant, set to 600
σ	A constant, set to 10
$\overline{Smax_n}$	Maximum sleep duration at NS n
MILP variables	
Ptx_{nt}	Adaptive transmission power of NS n at time t
$Ptxmin_{nt}$	Minimum transmission power of NS n at time t corresponding to maximum acceptable delay
ES_n	Energy saving of NS n
ESs_n	Energy saving of NS n with random sleep cycles
$ESra_n$	Energy saving of NS n with rate adaptation
PSs_n	Power saving of NS n with random sleep cycles
Pre_{nt}	The part of the available renewable power used by NS n at time t
$Dmin_n$	Minimum average piece delay at NS s (when data rate is $drmax_{nt}$)
dr_{nt}	Adaptive data rate at NS n at time t
$\overline{S_n}$	(Mean) Sleep duration at NS n
$\overline{Smax_n}$	Maximum (mean) sleep duration at NS n
NOs_n	Number of sleep cycles at NS n
Dra_n	Average piece delay at NS n with rate adaptation
Dsc_n	Average piece delay at NS n with sleep cycles
α_n	Equals 1 if NS n is ON, equals 0 otherwise
λ_{njt}	Traffic between NS n and TP j at time t
λ_n	Traffic demand at NS n at time t
δ_{njt}	Equals 1 if NS n is transmitting information piece to TP j, equals 0 otherwise
μmin_n	Minimum service rate at NS n
K_f	A variable to denote excess renewable energy compared to the maximum required
Index	
n	Index of Nano Servers (NSs)
j	Index of Traffic Point (TP)
t	Index of Time Point (T)

the vehicles (TP), which consumes negligible energy. We consider a scenario where the operational power consumption of the NS comprises of storage and CPU cycles for receiving requests from the vehicles.

The objective is to minimise the total energy consumption of the NSs over a given time. Hence, the objective function of the MILP model is

$$Minimize \quad \sum_{n \in NS} \alpha_n (Ptxmax_{nt} + Pidle_{nt}) \quad \forall n \in NS, \ \forall t \in T \tag{1}$$

$$Subject\ to \quad \lambda_{jt} = \sum_{n \in NS[j]} \lambda_{njt} \quad \forall j \in N[j], \ \forall t \in T \tag{2}$$

Equation (2) is the flow conservation constraint, which ensures that at each time point the total traffic is served. The control variable is α_n, it determines which Nano Server is to be turned ON.

$$\sum_{j \in tp[c]} \lambda_{njt} \le Bmax \quad \forall n \in NS, \ \forall t \in T. \tag{3}$$

Equation (3) ensures that the traffic demand at each NS does not exceed the capacity of a NS. If the NS has already reached its full capacity, the model should install another NS to serve the remaining traffic of the TP.

$$\frac{\lambda_{njt}}{\mu^{con}} \le A\delta_{njt} \tag{4}$$

where

$$\mu con = \frac{Bmax}{Ndmax} \quad \forall n \in NS, \ \forall j \in N[j], \ \forall t \in T \tag{5}$$

The constant A is a large positive number, used to enable the conversion from continuous variables to their binary equivalent. It is used to ensure that inequalities in (4) and (8) work. For example in (8), if the left hand side is nonzero, then this forces α_n to be

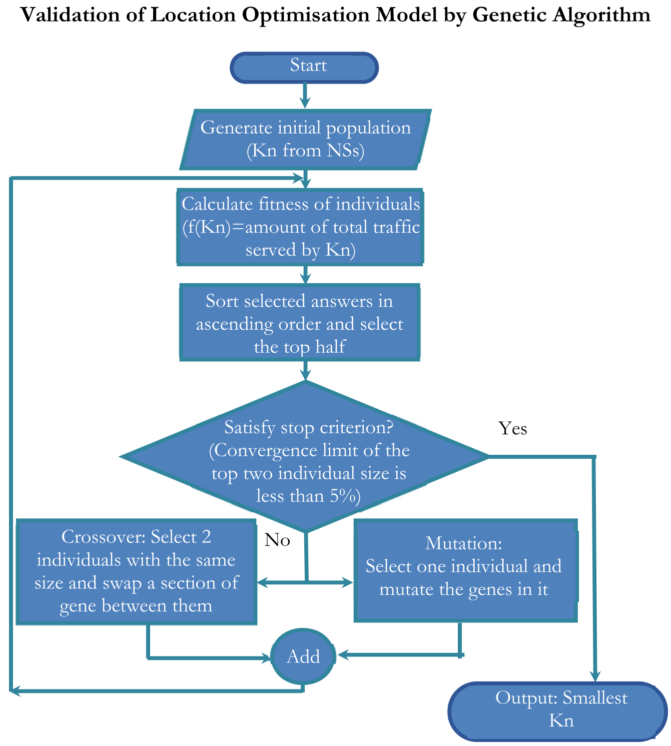
Validation of Location Optimisation Model by Genetic Algorithm
Start
Generate initial population (Kn from NSs)
Calculate fitness of individuals (f(Kn)=amount of total traffic served by Kn)
Sort selected answers in ascending order and select the top half
Satisfy stop criterion? (Convergence limit of the top two individual size is less than 5%)
Yes
No
Crossover: Select 2 individuals with the same size and swap a section of gene between them
Mutation: Select one individual and mutate the genes in it
Add
Output: Smallest Kn

Fig. 3. Location optimisation using genetic algorithm.

nonzero, for example, equal to one. This can only happen if the binary α_n is multiplied by a large positive number to ensure it is greater than the left hand side. Equation (4) ensures that if the traffic is non-zero between NS n and TP j, i.e. $\lambda_{njt} \neq 0$, then MILP Model 1 sets $\delta_{njt} = 1$. This is ensured by setting the constant A to a number, which depends upon the magnitude of the traffic between a TP and a NS. In the present case, $A = 600$ is used, which makes the right hand side comparable with the left hand side.

$$\frac{\lambda_{njt}}{\mu^{con}} \geq \delta_{njt} \quad \forall n \in NS,\ \forall j \in N[j],\ \forall t \in T. \tag{6}$$

Equation (6) ensures that if the amount of traffic between NS n and TP j is zero, i.e. $\lambda_{njt} = 0$, then there is no connection between them. Hence, MILP Model 1 sets $\delta_{njt} = 0$.

$$\sum_{t \in T} \sum_{j \in N[j]} \delta_{njt} \geq \alpha_n \quad \forall n \in NS \tag{7}$$

$$\sum_{t \in T} \sum_{j \in N[j]} \delta_{njt} \leq A\alpha_n \quad \forall n \in NS. \tag{8}$$

Equations (7) and (8) ensure that if there is a connection between NS n and TP j at time point t, then NS j is switched ON ($\alpha_n = 1$). The constant A is set to 600 which makes the right hand side (number of connections from different TPs to a NS) comparable with the left hand side (binary α_n).

The optimised locations and numbers of the NSs (Model 1) are validated using Genetic Algorithm (GA1) as shown in Fig. 3. Genetic Algorithms (GAs) are adaptive heuristic algorithms for finding the best possible solution with specific properties in a collection of competitive set of solutions based on the progressive ideas of natural selection and genetics. The genetic operators used in our algorithm are: Selection, Crossover, and Mutation.

The genetic algorithm validation results (Fig. 4) are very robust and validate our MILP model well (4% deviation). It is to be noted that the number of active NSs increases considerably during peak hours of the day, which is expected. Interestingly, the number of active NSs does not follow the variation of the total traffic because the number also depends upon the spatial distribution of the traffic. Thus, at some non-peak hours, a greater number of active NSs may be needed to serve the traffic demand. In each hour, a set of installed NSs remained switched off resulting in 74% reduction in daily energy consumption.

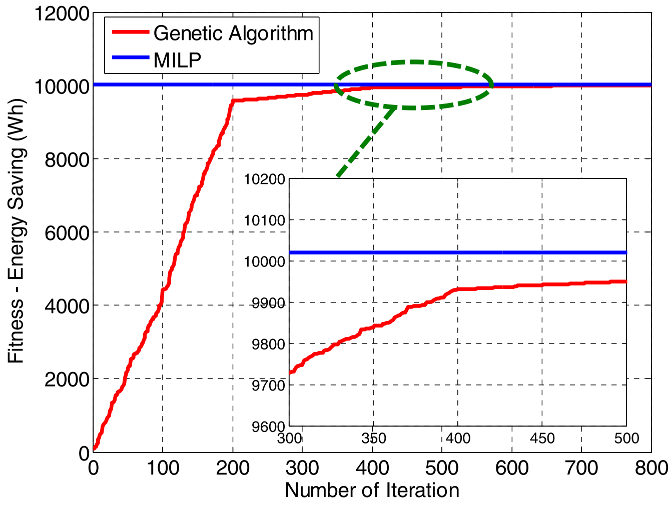

Fig. 4. Fitness value for different number of iterations.

5. Transmission energy minimisation of a Nano Server through random sleep cycles and rate adaptation

Location optimisation in the previous section was based on traditional NSs. In this section, we define two different techniques: (a) random sleep cycles (SC) and (b) rate adaptation (RA) for transmission energy minimisation of a NS. These give rise to reduced transmission energy consumption at the cost of increasing the average piece delay. In the remaining part of this section, we define the corresponding analytic models and validate them with simulation.

5.1. Analytic model of a Nano Server with multiple random sleep cycles

Here, the Nano Server operates random sleep cycles, which reduce the energy consumption by switching OFF only the transmitting circuitry part for a specific time duration randomly distributed with a mean value, when there is no request to serve. The RSU remained in sleep mode for a fixed time duration (randomly generated with a certain mean value) even if pieces were waiting to be served. Upon waking up, the NS served the arrived pieces (if any) and switched back to sleep mode when the buffer became empty. Each NS has a large buffer to hold the information piece requests. The arrival process of the Information piece requests is Poisson distributed. When the transmission (and corresponding processing) unit of the NS is idle, there is only idle power consumption ($Pidle$ in Table 1). Since the pieces are of fixed length, the service time is Deterministic. The sleep durations of random sleep cycles are random i.e. Negative exponential distributed with a certain mean value. By multiple random sleep cycles, we mean that the NS can switch to sleep mode multiple times [28]. Each sleep cycle (SC) is associated with a wake-up overhead (Ewo) [28]. The total energy overhead of a NS with multiple sleep cycles is dependent upon the number of times (NOs) the NS sleeps and wakes up within a given time duration.

In this section, the NS is modelled as an M/D/1/∞ queue with queue length dependent vacations. We term this as Model 2 in this paper. We solve the queueing model with residual life approach as illustrated in Fig. 5. The service duration of the NS is deterministic with mean, $\overline{X} = \mu^{-1}$ and variance $Var(X) = 0$. Furthermore,

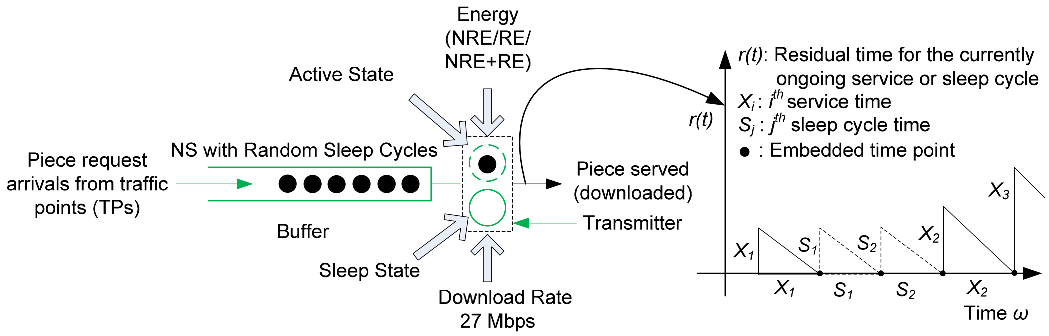
Energy (NRE/RE/ NRE+RE)
Active State
NS with Random Sleep Cycles
Piece request arrivals from traffic points (TPs)
Buffer
Sleep State
Download Rate 27 Mbps
Piece served (downloaded)
Transmitter
r(t)
r(t): Residual time for the currently ongoing service or sleep cycle
X_i: i^{th} service time
S_j : j^{th} sleep cycle time
● : Embedded time point
X_1
S_1
S_2
X_2
X_3
X_1
S_1
S_2
X_2
Time ω

Fig. 5. M/D/1/∞ queues with sleep cycles to represent a NS operation.

the sleep cycles are Negative exponential distributed with mean $\bar{S}$. The service durations and sleep cycle durations are independent and identically distributed (i.i.d.) random variables, which are also independent of each other.

Let $r(\omega)$ be the residual time for the ongoing service or sleep. The time average of $r(\omega)$ $(0 \le \omega \le t)$ can be given as

$$\overline{R_t} = \frac{1}{t}\int_0^t r(\omega)d\omega = \frac{1}{t}\sum_{i=1}^{N_\lambda(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{j=1}^{NOs(t)}\frac{1}{2}S_j^2 \tag{9}$$

where $NOs(t)$ is the number of information piece requests that arrive within $(0, t)$, and ω is the instantaneous time.

$NOs(t)$ is the number of sleep cycles (sleep count) within $(0, t)$. Therefore,

$$\begin{aligned}\overline{R} &= \lim_{t\to\infty}\overline{R_t} = \lim_{t\to\infty}\left[\frac{1}{t}\sum_{i=1}^{N\lambda(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{j=1}^{NOs(t)}\frac{1}{2}S_j^2\right]\\ &= \lim_{t\to\infty}\left[\frac{N\lambda(t)}{t}\frac{\sum_{i=1}^{N\lambda(t)}\frac{1}{2}X_i^2}{N\lambda(t)}\right] + \lim_{t\to\infty}\left[\frac{NOs(t)}{t}\frac{\sum_{j=1}^{NOs(t)}\frac{1}{2}S_j^2}{NOs(t)}\right]\\ &= \frac{1}{2}\lambda\overline{X^2} + \frac{1}{2}\lim_{t\to\infty}\frac{1}{\frac{t(1-\rho)}{NOs(t)}}\overline{S^2}(1-\rho)\\ &= \frac{1}{2}\lambda\overline{X^2} + \frac{1}{2}(1-\rho)\frac{\overline{S^2}}{\overline{S}}\end{aligned} \tag{10}$$

where the offered load ρ also corresponds to system utilisation (Us) and is defined as

$$\rho = Us = \lambda\overline{X}. \tag{11}$$

The relation between ρ, NOs and $\bar{S}$ can be defined as

$$\frac{t(1-\rho)}{NOs} = \bar{S} \tag{12}$$

where t represents a time duration.

The average waiting time for an information piece at the NS can be computed using Little's theorem [42] as

$$\begin{aligned}Wq &= NOq\overline{X} + \overline{R}\\ &= \lambda Wq\overline{X} + \overline{R}\end{aligned} \tag{13}$$

where NOq represents the number of information piece requests waiting in the queue.

Therefore,

$$Wq = \frac{\overline{R}}{(1-\rho)} \tag{14}$$

Since $\rho = \lambda\overline{X}$, using Eq. (10) and Eq. (11) we get

$$Wq = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\overline{S^2}}{2\overline{S}} \tag{15}$$

The average delay of an information piece (Dsc) including the waiting delay (Wq) and the service time $(\overline{X})$ can be computed as

$$\begin{aligned}Dsc &= Wq + \overline{X}\\ &= \overline{X} + \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\overline{S^2}}{2\overline{S}}\end{aligned}$$

Since $\overline{X} = \frac{1}{\mu}$, $\overline{X^2} = Var(X) + (\overline{X})^2 = \frac{1}{\mu^2}$ where μ is the service rate (number of pieces per second) $(Var(X) = 0$ as data rate is deterministic)

$$\begin{aligned}&= \frac{1}{\mu} + \frac{\lambda}{2\mu^2(1-\rho)} + \frac{\overline{S^2}}{2\overline{S}}\\ &= \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)} + \overline{S}\end{aligned} \tag{16}$$

Since $\overline{S^2} = Var(S) + (\overline{S})^2 = (\overline{S})^2 + (\overline{S})^2 = 2(\overline{S})^2$. $(Var(S) = (\overline{S})^2$ as sleep cycle is random.)

The average number of pieces in the system $(Nsys)$ from Little's theorem [42] can be written as

$$\begin{aligned}Nsys &= \lambda Dsc\\ Nsys &= \lambda\overline{X} + \frac{\lambda^2}{2}\frac{\overline{X^2}}{(1-\rho)} + \frac{\lambda}{2}\frac{\overline{S^2}}{\overline{S}}\\ &= \rho + \frac{\rho^2}{2(1-\rho)} + \overline{S}\end{aligned} \tag{17}$$

The energy savings (Es) per hour through sleep cycles for the NS can be expressed as

$$\begin{aligned}Es &= (1-Us)\cdot Ptxmax\cdot T - (Ewo\cdot NOs)\\ &= (1-\rho)\cdot Ptxmax\cdot T - (Ewo\cdot NOs)\\ &= \frac{NOs\cdot\overline{S}}{T}\cdot Ptxmax\cdot T - (Ewo\cdot NOs)\\ Es &= NOs(\overline{S}Ptxmax - Ewo)\end{aligned} \tag{18}$$

where Us is the system utilisation, $Ptxmax$ is the NS's transmitter circuitry power consumption and Ewo is the NS's wake-up overhead. The number of sleep cycles i.e. sleep count per unit time $(NOs(t))$ can be computed from (12) as

$$\lim_{t\to\infty}\frac{NOs(t)}{t} = NOs = \frac{(1-\rho)}{\overline{S}} \tag{19}$$

Rearranging (16), in terms of $\overline{S}$, we obtain

$$\overline{S} = Dsc - \frac{2\mu-\lambda}{2\mu(\mu-\lambda)} \tag{20}$$

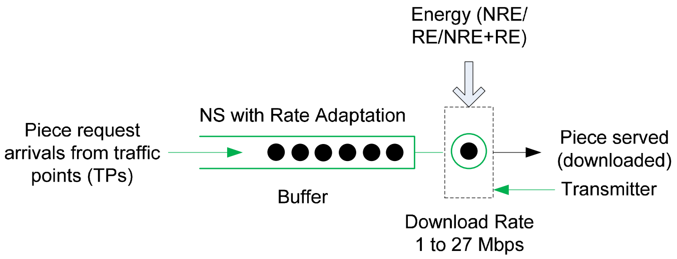

Fig. 6. M/D/1/∞ queue with rate adaptation, used to represent NS operation.

The MILP constraint for transmission energy minimisation through sleep cycles can therefore be written below.

When the maximum delay is allowable with maximum service rate i.e. $Dsc = Dmax$ and $\mu = \mu max$, the condition for maximum energy savings at nth NS is

$$\bar{S}_n max = Dmax - \frac{2\mu max - \lambda_n}{2\mu max(\mu max - \lambda_n)} \quad \forall n \in NS \tag{21}$$

$$NOs_n = \frac{(1 - \frac{\lambda_n}{\mu max})}{\bar{S}_n max} \quad \forall n \in NS \tag{22}$$

$$Esmax_n = NOs_n(\bar{S}_n max Ptx_n - Ewo) \quad \forall n \in NS \tag{23}$$

The representation of the information piece delivery through a suitable model was necessary in order to derive performance equations, which are subsequently used in the MILP. Although the queueing discipline M/D/1 is straight forward, however, deriving those limiting conditions and relationship among the performance parameters are the new developments in this work. To elaborate, we can find that Eqs. (21), (22), (23) form the basis of transmission energy minimisation through sleep cycles. Equation (21) defines the relationship between maximum sleep duration and maximum information piece delay for a given traffic. Equation (22) defines the relationship between the number of sleep cycles (sleep count) and maximum sleep duration for a given traffic. Equation (23) defines the relationship between maximum energy savings, sleep count, maximum sleep duration and wakeup overhead.

5.2. Analytic model of a NS with rate adaptation

For the rate adaptive NSs, the service rate adapts according to the energy available for transmission. In our case the service rate of a NS varies linearly between 3 Mbps and 27 Mbps. Assuming that the service rate is linearly proportional to the transmission energy, the adaptive service rate of a NS (μ) can be expressed in terms of transmission energy (Ptx) as

$$\mu = \left(\frac{Ptx}{Ptxmax}\right)\mu max \tag{24}$$

where μ is the non adaptive (maximum) service rate and $Ptxmax$ is the maximum transmitting power (e.g. 7 W). Invariably, $Ptxmin \le Ptx \le Ptxmax$ and $\mu min \le \mu \le \mu max$. Thus, the transmission energy consumption of a NS reduces compared to that of a non-rate adaptive NS. This, in turn, further reduces the total energy consumption.

To find out the relationship between the service rate at each NS and the average information piece delay, we analyse the stochastic properties of the arrival and the departure processes of a NS. The arrival process of information piece requests at any NS is random and can be modelled as a Poisson process. The information piece size is fixed (deterministic) and the service rate is also fixed for a fixed supply of transmission energy. In our case, the transmission energy is wind which varies with a resolution of 15 minutes (does not change within a short time). Thus, the wind energy imparts constant power for transmission purposes and therefore a constant service rate for an information piece. Thus, the service time of an information piece is deterministic. Such a NS can be modelled as a M/D/1 queue (Model 3) (Fig. 6). The average information piece delay in a rate adaptive NS can be determined by setting $\bar{S} = 0$ in (16) and rearranging, which gives

$$Dra = \frac{2\mu - \lambda}{2\mu^2 - 2\mu\lambda}. \tag{25}$$

The MILP constraint for transmission energy minimisation through rate adaptation can therefore be written as: $Dra = Dmax$, $\mu = \mu min$, which is the condition for maximum energy savings through rate adaptation at a NS. Solving (25), we obtained the minimum service rate for nth NS as

$$\mu min_n = \frac{2\lambda_n Dmax + 2 \pm \sqrt{4\lambda_n^2(Dmax)^2 + 4}}{4Dmax} \quad \forall n \in NS \tag{26}$$

The minimum power consumption of a rate-adaptive NS can be expressed as

$$Pmin_n = Ptxmin_n + Pidle \quad \forall n \in NS \tag{27}$$

where $Pmin_n$ is the minimum power required by NS n for serving its traffic with the maximum acceptable information piece delay. The quantity $Pmin_n$ is computed by expressing (25) as:

$$\sum_{t \in T} \sum_{j \in tp[n]} \delta_{njt} \neq 0 \quad \forall n \in NS \tag{28}$$

$$\begin{aligned} &\mu RA_n \\ &\quad = \frac{\sum_{t \in T} \sum_{j \in N[n]} \delta_{njt}.Dra_n + 1}{2Dra_n} \\ &\qquad + \frac{\sqrt{(\sum_{t \in T} \sum_{j \in tp[n]} \delta_{njt}.Dra_n + 1)^2 - 2\sum_{t \in T} \sum_{j \in tp[n]} \delta_{njt}.Dra_n}}{2Dra_n} \\ &\quad \forall n \in NS. \end{aligned} \tag{29}$$

For minimum adaptive service rate (μra_n), we write

$$Dra_n = Dmax = 13s \quad \forall n \in NS \tag{30}$$

$$Ptxmin = \left(\frac{\mu min}{\mu max}\right) Ptxmax \quad \forall n \in NS. \tag{31}$$

When the energy source is non-renewable, the service rate of a rate adaptive NS reaches its minimum value needed to maintain the average piece delay below the acceptable limit (13 s).

The formulation of rate adaptation is very simple. The idea is to use minimum energy (transmission rate) that maintains an acceptable delay for the information piece delivery through a suitable queueing model, which is subsequently used in the MILP. Although the queueing discipline M/D/1 is straight forward, however, deriving those limiting conditions and relationships among the performance parameters are the new developments in this work. To elaborate, we can find that Eqs. (21), (22), (23) forms the basis of transmission energy minimisation through sleep cycles. Equation (21) defines the relationship between maximum sleep duration and maximum information piece delay for a given traffic. Equation (22) defines the relationship between the number of sleep cycles (sleep count) and maximum sleep duration for a given traffic. Equation (23) defines the relationship between the maximum energy savings, sleep count, maximum sleep duration and wakeup overhead.

5.3. Model validation of a NS with sleep cycles and rate adaptation

To validate the models, we obtain the energy savings from Model 2 and Model 3 for a fixed average piece delay through analytic modelling and simulation. For analytic results, the corresponding expressions for energy savings are given below.

From Model 2, we obtain a relationship between energy savings (Es_{nt}) and mean arrival rate of piece requests (λ) in terms of fixed average piece delay ($D_1 = 5s$) using (18), (22) and (23) as

$$Essc_{nt} = Ptxmax_{nt} \cdot \left(1 - \frac{\lambda}{\mu}\right) \cdot T - \left(\frac{(\mu-\lambda)^2}{2\mu^2 D_1 - 2\mu(\lambda D_1 + 1) + \lambda}\right) Ewo \tag{32}$$

Similarly from Model 3, we obtain a relationship between energy savings per hour (Es) and mean arrival rate of piece requests (λ) in terms of fixed average piece delay ($D_2 = 5s$) using (24), (25), (26) and (30) as

$$Esra_{nt} = Ptxmax_{nt} \cdot \left(1 - \frac{Dmin \cdot (2\lambda D_2 + 2 \pm \sqrt{4\lambda^2 D_2^2 + 4})}{D_2 \cdot (2\lambda Dmin + 2 \pm \sqrt{4\lambda^2 Dmin + 4})}\right) \tag{33}$$

For simulation, we developed a rate adaptive algorithm and sleep cycles algorithm for a NS, which are described as follows. The rate adaptation algorithm determines the data rate from the available energy to the NS. It then records the delay of the individual pieces served. Finally, it computes the average delay of the served pieces. The algorithm determines the data rate from the available energy to the NS. It then records the delay of the individual pieces served. Finally, it computes the average delay of the served pieces. The sleep cycles algorithm determines the sleep duration from the state of the transmitter of the NS each time there is no request to serve. It then determines the average piece delay.

We have also found by analysis and simulation that the energy savings achieved by Model 2 (NS with sleep cycles) and Model 3 (Rate adaptive NS) decrease with increase in the arrival rate of piece requests for a fixed average piece delay (5s). Since, there is no reference wake-up overhead value for a NS available in the literature, we assume two cases where the wake-up process consumes 10% and 20% of the maximum transmission power of a NS. It is evident that the higher the wake-up overhead (20%) the lower energy savings for the same load. With rate adaptation, the service rate of a NS increases with increase in load. This results in decrease in energy savings.

5.4. MILP model for optimum usage of non-renewable energy and renewable energy mix

In this section, we utilise the renewable (wind) energy (RE) available from wind farm to reduce the non-renewable energy consumption of the NSs. The corresponding MILP models (4 and 5) minimise non-renewable energy (NRE) consumption subjected to the (maximum) average piece delay constraint and subsequently improve the delay with the available RE. The models ensure that heavily loaded NSs receive greater RE to improve the overall user experience. MILP model 4 uses NSs with sleep cycles and MILP model 5 uses rate adaptive NSs.

5.4.1. Sleep cycle

Model 4 is developed based upon MILP Model 1. The optimised NSs location is used as input here. Therefore, the objective function of the MILP Model 4 is to minimise the power consumption by maximising the energy savings ($\forall t \in T, \alpha_n = 1$). NOs_n is not a variable in this MILP. The number of sleep cycles depends upon the packet arrival statistics at each NS. This is dictated by the average piece requests (λ), computed from the location optimisation model.

$$\text{Minimize} \quad \sum_{t \in T} \sum_{n \in NS} \alpha_n (Ptxmax_{nt} + Pidle_{nt} - Ps_n - Pre_{nt})$$
$$\text{where } Ps_n = Ess_n / T \tag{34}$$

and α_n is a constant to make the optimisation work which is duly adjusted for each Nano Server. The constraints of the above MILP (model 4) are given below.

$$\overline{S_n} \leq \overline{Smax_n} \quad \forall n \in NS \tag{35}$$

$$Pre_{nt} \leq Ptxmax_{nt} + Pidle_{nt} - Psmax_n^s \quad \forall n \in NS, \ \forall t \in T \tag{36}$$

where (from Eq. (23)),

$$Psmax_n^s = NOs(\overline{Smax_n} Ptx_{nt} - Ewo)/T \quad \forall n \in NS, \ \forall t \in T \tag{37}$$

$$PRE_n = Pw_n - K_n \quad \forall n \in NS \tag{38}$$

$$Ptxmax_n + Pidle_n - Psmax_n - PRE_n \geq 0 \quad \forall n \in NS \tag{39}$$

Equations (37) and (38) ensure that the NRE consumption by NS n is non-negative, i.e. $Pre_n \geq Ptxmax_n + Pidle_n - Psmax_n$. Hence, K_f is the difference between operational energy consumption of NS n and the amount of available renewable energy.

The constraint for sleep duration is

$$\overline{S_n} = Dsc_n - \frac{2\mu \max_n - \lambda_n}{2\mu \max_n (2\mu \max_n - \lambda_n)} \quad \forall n \in NS \tag{40}$$

The sleep duration $\overline{S_n}$ of a NS is tuneable based upon the available renewable energy. The demand (λ_n) at each NS is computed in MILP Model 1.

With excess RE, the MILP model 4 reduces sleep duration at busy NSs, which experience higher piece delay. This is achieved by the constraint

$$\overline{S_k} - \overline{S_i} \geq \sigma(\lambda_i - \lambda_k) \quad \forall i \in NS, \ k \in NS, \ i \neq k \tag{41}$$

where σ is a constant which makes the right hand side (traffic demand) comparable with the left hand side (sleep duration).

5.4.2. Rate adaptation

The objective function of the MILP model 5 is

$$\text{Minimize} \quad \sum_{t \in T} \sum_{n \in NS} \alpha_n (Ptx_{nt} + Pidle_{nt} - Pre_{nt})$$
$$\forall n \in NS, \ \forall t \in T \tag{42}$$

The constraints of the MILP model 5 are given below.

$$Dra_n \leq Dmax_n \quad \forall n \in NS \tag{43}$$

The minimum required energy ($Pmin_{nt}$) is calculated using (26) where the service/data rate of a NS is adaptive based on the available renewable energy for transmission. Equation (35) is modified as below.

$$Pre_{nt} \leq Ptx_{nt} + Pidle_{nt} \quad \forall n \in NS, \ \forall t \in T \tag{44}$$

The transmission energy constraint is given below:

$$Ptx_{nt} + Pidle_{nt} \geq Pmin_{nt} \quad \forall n \in NS, \ \forall t \in T \tag{45}$$

The constraint for data rate adaptation is given below

$$dr_{nt} = drmax_{nt} \frac{Ptx_{nt}}{Ptxmax_{nt}} \quad \forall n \in NS, \ \forall t \in T \tag{46}$$

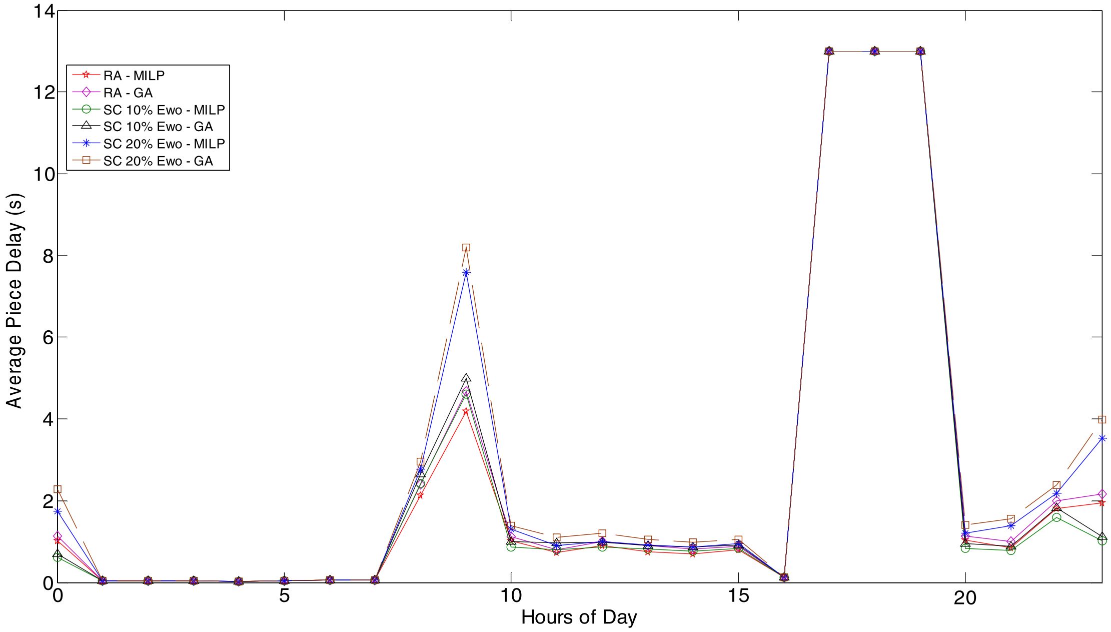

Fig. 7. Average piece delay validation.

Equation (46) computes the adaptive data rate of each NS at each time point t. To compute the adaptive transmission power of Sn, if it has a connection with any $TP(\alpha_n = 1)$, we need to linearise Eq. (46). Hence, we define X_{nt} as

$$X_{nt} = Ptx_{nt}\alpha_n \quad \forall n \in NS, \ \forall t \in T \tag{47}$$

The following Eqs. (48)–(50) are used to remove the non-linearity which results from multiplying two variables, and replace the relations with equivalent linear relationship.

$$0 \le X_{nt} \le Ptxmax_{nt} \quad \forall n \in NS, \ \forall t \in T \tag{48}$$

$$X_{nt} \le \alpha_n Ptxmax_{nt} \quad \forall n \in NS, \ \forall t \in T \tag{49}$$

$$X_{nt} \ge Ptx_{nt} - Ptxmax_{nt}(1 - \alpha_n) \quad \forall n \in NS, \ \forall t \in T. \tag{50}$$

With the excess RE, the proposed model improves data rate at busy NSs, which experienced lower data rate. This is achieved by the constraint

$$dr_{it} - dr_{kt} \ge \sigma(\lambda_i - \lambda_k) \quad \forall i\&k \in NS, \ i \ne k \tag{51}$$

where σ is a constant which makes the right hand side (traffic demand) comparable with the left hand side (data rate).

5.4.3. Validation of optimum usage of non-renewable energy and renewable energy mix model using genetic algorithms

We develop another genetic algorithm (GA2) for validating Model 4 and Model 5, which determines the optimum mixture of renewable and non-renewable energy based on load demand profile. In GA2, we start working with a population of individuals (corresponds to a random amount of renewable energy for each TP) having any “fitness” (average piece delay) and then allow that population to evolve to a more fit state. The better-fit individuals having lower delay are more likely to mate than an individual that is poorly fit to survive in the new environment. Thus, the genes of the lower delay solutions are more likely to be passed on to the offspring.

The GA begins by generating the initial population randomly, which in the present case is a set of 400 locations of the NSs. Each pair of chromosomes is of equal size but the sizes are random between 1 and 398. The next step is the selection operation. The top half chromosomes from the population based on the amount of traffic, which they can serve are selected to pass on their genes to the next generation. Crossover needs to be done on these selected chromosomes and one offspring is kept to add to the top selected population. Mutation needs to be done and is applied to 10% of the population. Mutation in our genetic algorithm is random in one gene, where the mutated chromosomes replace the existing chromosomes. Thus, the population size remains the same. In every iteration, the chromosomes are sorted by the amount of traffic they can serve until the stop criterion is satisfied. The GA checks the size of each chromosome and outputs the chromosome with the smallest cost to minimise power consumption.

Fig. 7 shows the average piece delay achieved at each hour for (i) the NSs with random sleep cycles having different wake-up energy overhead and (ii) the rate adaptive NSs through MILP and Genetic Algorithms. The GA result deviates 5.9% compared to that of MILP for RA and 6.1% and 6.5% compared with that of SC (10% wake up overhead) and SCs (20% wakeup overhead), respectively. This algorithm is similar to GA1 apart from the initial population size and fitness factor. Chromosomes have renewable power allocations as array elements and the size of each array is equal to the total number of NSs at each hour. Renewable power is related to delay in case of sleep cycles through Eqs. (18) and (23), and in case of rate adaptation through Eqs. (25) and (26), respectively. It can be concluded that GA is a very robust and good tool for validating the MILP model and can work as an alternative to MILP in the related problem scenario.

6. Results and discussion

In this section, the performance of the proposed NSs is evaluated in terms of energy consumption and average piece delay.

Fig. 8 shows the locations of the Traffic Points (TPs) which represent the traffic centroids that account for the traffic (information piece/data) which the nearby vehicles request and the locations of active NSs for whole day. By installed NSs, we mean those CS where NSs need to be active at least once in the whole day. The locations of the installed NSs can be obtained by the union of the hourly sets of active NSs over the whole day. At each hour, a subset of the NSs are active among the entire set of installed NSs, while the others are switched off. The MILP Model 1 reduces the

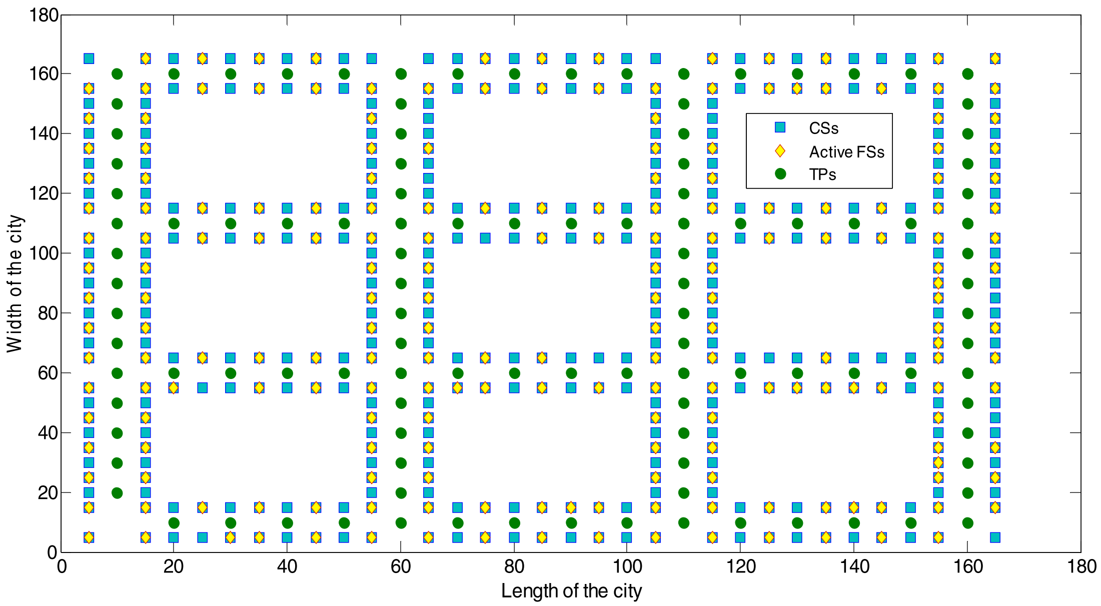

Fig. 8. Locations of the CSs, TPs and installed NSs.

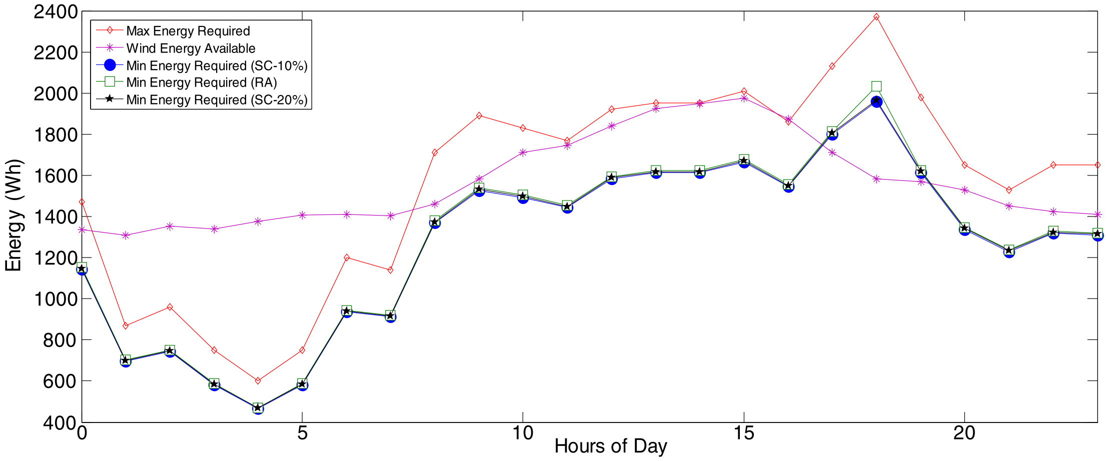

Fig. 9. Available and required energy for the rate-adaptive NSs.

required number of installed NSs from 398 to 202 resulting in 49% reduction in overall energy consumption of the NSs. The chosen 3 km × 3 km area represents a typical city centre zone. While it is of interest to evaluate larger inner city zones, the constrained mixed integer linear programming optimisation is at the limit of our computational capabilities for this size of inner city. We have employed a high performance computing cluster with 16 cores and 256 GB RAM in the solution. We plan in future to extend the work by introducing fast solutions to the sparse linear MILP equations to enable us to tackle larger inner city problems.

Fig. 9, shows the available wind energy from the wind farm, the minimum and the maximum required energies by the installed NSs. By minimum required energy, we mean the least amount required to maintain the average piece delay under the mentioned bound. The relationship is obtained through the queueing model in Section 4. At early hours of the day, the available wind energy is considerably higher compared to the required energies. However, it becomes comparable during peak hours due to higher traffic load at those periods. Since the available wind energy and the required energy (minimum or maximum) are independent of each other, the available energy is even lower than the minimum required energy at some instances, i.e. hours 17–19. Consequently, these hours require NRE to meet the deficiency.

The authors in [43] analysed the performance of rate adaptation techniques based on the concept of 'coherence time' using a channel emulator. Moreover, a rate selection policy was presented based on the speed and location of a vehicle [31]. The rate adaptation concept can also be used for energy efficiency where the service quality varies according to the available energy. This can be termed: Energy aware rate adaptation.

The optimal energy consumptions of the rate adaptive NSs and NSs with Sleep Cycles are shown in Fig. 10, when an optimal mix of NRE and RE are used (through MILP Models 2 and 3). As can be seen, apart from a few occasions where available wind energy was lower than that of minimum required energy, the available wind energy on its own was sufficient to operate the NSs while improving the service rate.

As mentioned earlier, the maximum energy consumption corresponds to the consumption of installed NSs without rate-adaptation. The non-rate adaptive (traditional) NSs operate at maximum service rate giving rise to the lowest piece delay. Similarly, minimum energy consumption refers to the NRE consumption of the rate adaptive NSs operating at minimum service rates. Such service rates are bounded by the maximum average piece delay requirement for respective traffic loads. The optimal energy consumption refers to the situation, where the MILP Model 4 selects

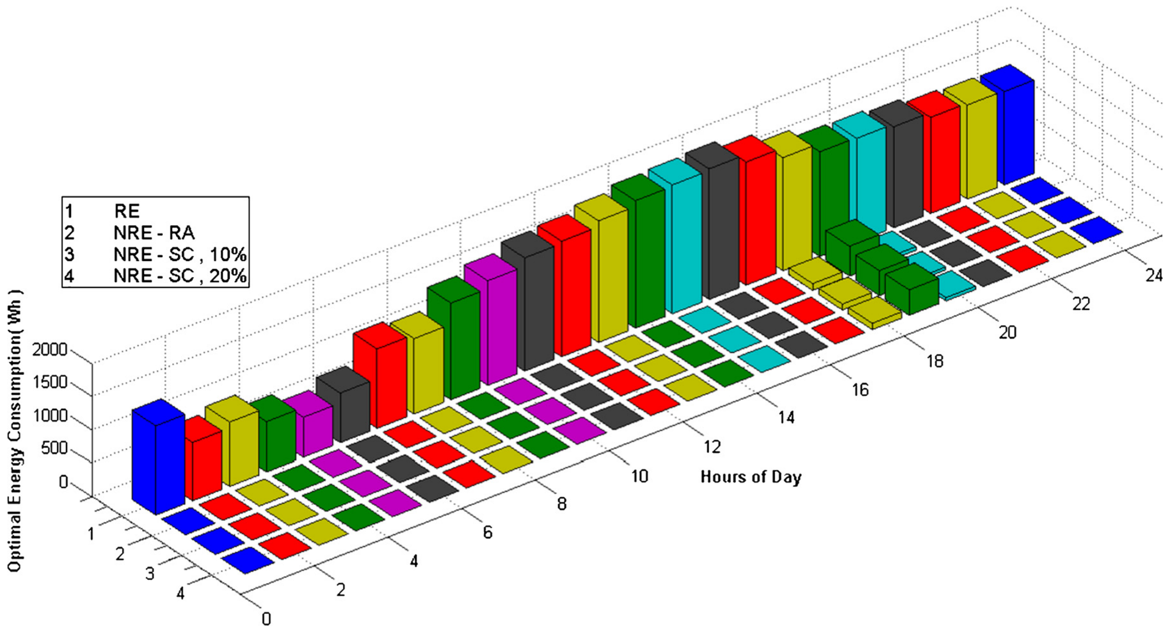

Fig. 10. Optimal energy consumption of the rate-adaptive NSs and NSs with Sleep Cycles.

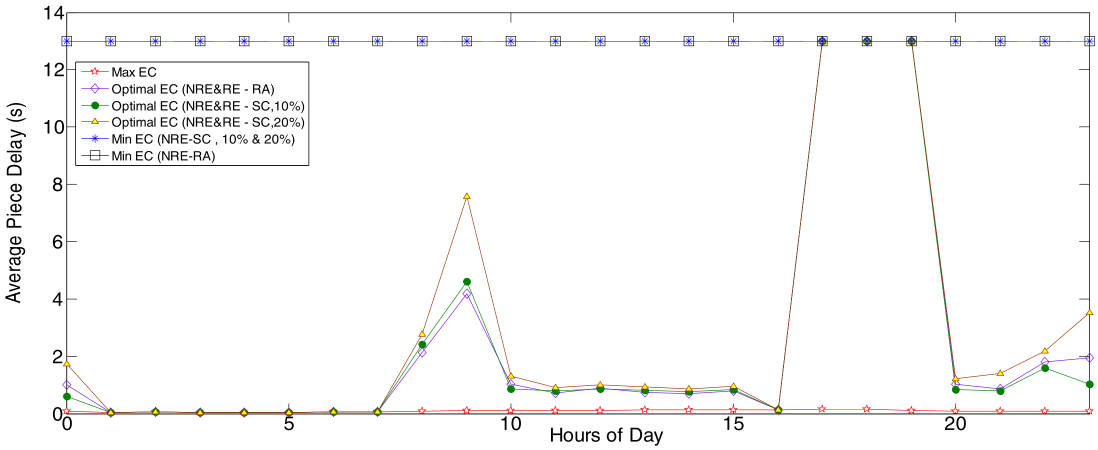

Fig. 11. Average piece delay. “EC” stands for Energy Consumption.

the busy NSs with lowest service rates and tries to maximise the service rates with the available excess RE. This effectively minimises the carbon footprint and also yields maximum improvement in the achieved service rate. For the NSs with random sleep cycles, maximum energy is consumed when no sleep takes place and minimum energy is consumed when sleep cycles occur with maximum duration. Similarly, optimal energy consumption refers to the load dependent renewable energy distribution, where the NSs operate with variable sleep duration to decrease the waiting delay. Thus, the average piece delay significantly reduces during most of the day (Fig. 11).

7. Conclusions

In this paper, we proposed sleep enabled and rate adaptive Nano Servers for energy efficient delivery of information pieces in a smart city vehicular environment. A rate adaptive NS operates at variable service rate according to the available transmission energy while sleep enabled NS switched to sleep mode to save transmission energy. Firstly, the proposed MILP Model 1 optimised the number of (traditional non-rate adaptive) NSs for a typical smart city scenario through a continuous selection process of NSs (transient optimisation) according to the varying hourly traffic. The number and location optimisation of NSs resulted in 49% energy savings compared to a non-optimised setup. A further 74% energy saving was obtained through transient optimisation in which, the MILP model switched off the non-active NSs selectively based upon the spatial-temporal variation of the traffic.

Model 2 corresponds to NSs with sleep cycles, which saved up to 81% and 80% energy respectively for wake up overhead energy of 10% and 20% of the maximum transmission energy respectively. Similarly, Model 3 incorporated rate adaptive NSs, which saved up to 78% transmission energy without exceeding the average piece delay limit. The above savings in non-renewable energy were complemented by renewable energy available from a wind farm. The challenge of using renewable energy lied in its judicious distribution to the rate adaptive NSs such that the overall QoS can be maximised. With that aim in mind, we proposed MILP Models 4 and 5, which first replaced the non-renewable energy with the available renewable energy. Models 4 and 5 then distributed the excess renewable energy to improve QoS. The load dependent distribution of the excess renewable energy improved the individual service rate of the NSs in a way that the majority of the piece downloads experience minimum average piece delay. The MILP Model 4 improved the average piece delay by 82% and minimised the carbon footprint by 96% in the information piece delivery scenario in our smart city vehicular environment. Such a rate adaptive NSs is essential for future Internet of Things. The MILP Model 5 improved the average piece delay by 81% and 83% for corresponding wake up overhead energies of 10% and 20% of the maximum transmission energy, and minimised the carbon footprint by 98% in the same scenario. All the MILP models are validated by genetic algorithms, which are very efficient and produce results that are comparable to MILP. For the case of location optimisation, the deviation between the MILP and the genetic algorithm is 4%, while for the case of optimum usage of non-renewable energy and renewable energy mix, the deviation is about 6% (for both rate adaptation and sleep cycles).

Finally, it is to be noted that both sleep cycles and rate adaptation are effective with smaller devices like Nano Servers where the switching overhead is acceptable rather than in cloud architecture. Rate adaptation is further desirable as it does not incur switching overhead and thereby maintains continuous service. The results in this paper for sleep cycles are shown with an estimated energy wakeup overhead for the Nano Servers. However, the model is generic and can be used with any wakeup overhead. In future, with the implementation of technology, actual wakeup overhead would be available, which may alter the results case by case. In case of rate adaptation, performance is bounded by the minimum and the maximum service rates. Future directions in this topic may involve heterogeneous content delivery including local information, security and safety related messages, which might lead to complex optimisation problems.

Acknowledgements

The authors would like to acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC), UK for funding the INTelligent Energy aware NETworks (INTERNET) project under contract EP/H040536/1. All data are provided in full in the results section of this paper.

References

[1] J. Oueis, E.C. Strinati, S. Sardellitti, S. Barbarossa, Small cell clustering for efficient distributed fog computing: a multi-user case, in: IEEE 82nd Vehicular Technology Conference, VTC2015-Fall, Boston, MA, 2015, pp. 1–5.

[2] M. Li, Z. Yang, W. Lou, CodeOn: cooperative popular content distribution for vehicular networks using symbol level network coding, IEEE J. Sel. Areas Commun. 29 (January 2011) 223–235.

[3] D. Zhang, C.K. Yeo, Enabling efficient wifi-based vehicular content distribution, IEEE Trans. Parallel Distrib. Syst. 24 (March 2013) 479–492.

[4] U. Shevade, Y.-C. Chen, L. Qiu, Y. Zhang, V. Chandar, M.K. Han, et al., Enabling high-bandwidth vehicular content distribution, in: Proceedings of the 6th International Conference on Emerging Networking Experiments and Technologies, CoNEXT, Philadelphia, Pennsylvania, 2010.

[5] M. Shojafar, N. Cordeschi, E. Baccarelli, Energy-efficient adaptive resource management for real-time vehicular cloud services, IEEE Trans. Cloud Comput. 4 (2016) 1–14.

[6] P. Kolios, V. Friderikos, K. Papadaki, Ultra low energy store–carry and forward relaying within the cell, in: IEEE 70th Vehicular Technology Conference Fall, 2009, pp. 1–5.

[7] W. Kumar, S. Bhattacharya, B.R. Qazi, J.M.H. Elmirghani, An energy efficient double cluster head routing scheme for motorway vehicular networks, in: IEEE International Conference on Communications, ICC, 2012, pp. 141–146.

[8] W. Kumar, S. Bhattacharya, B.R. Qazi, J.M.H. Elmirghani, A vacation-based performance analysis of an energy-efficient motorway vehicular communication system, IEEE Trans. Veh. Technol. 63 (2014) 1827–1842.

[9] C. Sun, Y.D. Alemseged, H.N. Tran, H. Harada, Transmit power control for cognitive radio over a Rayleigh fading channel, IEEE Trans. Veh. Technol. 59 (2010) 1847–1857.

[10] W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks, in: IEEE Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 2000, p. 10.

[11] S. Igder, H. Idjmayyel, B.R. Qazi, S. Bhattacharya, J.M.H. Elmirghani, Load adaptive caching points for a content distribution network, in: 9th International Conference on Next Generation Mobile Applications, Services and Technologies, 2015, pp. 150–155.

[12] M. Boldi, S. Petersson, M. Fodrini, A. Orlando, P. Persson, A. Nilsson, Multi antenna techniques to improve energy efficiency in LTE radio access network, in: IEEE Future Network & Mobile Summit, 2011, pp. 1–8.

[13] R. Hussain, J. Son, H. Eun, S. Kim, H. Oh, Rethinking vehicular communications: merging VANET with cloud computing, in: 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, 2012, pp. 606–609.

[14] M.K. Sharma, A. Kaur, A survey on vehicular cloud computing and its security, in: IEEE 1st International Conference on Next Generation Computing Technologies, NGCT, 2015, pp. 67–71.

[15] TROPIC: Distributed computing, storage and radio resource allocation over cooperative femtocells, 7th EU Framework Programme (ed).

[16] F.R.M. Bonomi, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, 2012, pp. 13–16.

[17] N. Cordeschi, D. Amendola, M. Shojafar, E. Baccarelli, Distributed and adaptive resource management in cloud-assisted cognitive radio vehicular networks with hard reliability guarantees, Veh. Commun. 2 (2015) 1–12.

[18] C. Borcea, X. Ding, N. Gehani, R. Curtmola, M.A. Khan, H. Debnath, Avatar: mobile distributed computing in the cloud, in: 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, San Francisco, CA, 2015, pp. 151–156.

[19] T.S. Rappaport, S. Sun, R. Mayzusi, H. Zhao, Y. Azar, K. Wang, et al., Millimeter Wave Mobile Communications for 5G Cellular: It Will Work! IEEE Access 1 (2013) 335–349.

[20] S. Sen, J. Rexford, D. Towsley, Proxy prefix caching for multimedia streams, in: Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 1999, pp. 1310–1319.

[21] M. Johnson, L.D. Nardis, K. Ramchandran, Collaborative content distribution for vehicular ad hoc networks, in: Allerton Conference on Communication, Control, and Computing, 2006.

[22] C. Peng, C. Chen-Nee, L. Xin, Energy-aware node placement in wireless sensor networks, in: IEEE Global Telecommunications Conference, GLOBECOM, 2004, pp. 3210–3214.

[23] E. Callaway, Low power consumption features of the IEEE 802.15. 4/ZigBee IR-WPAN standard. Mini-tutorial, in: ACM Sensys, vol. 3, 2003, pp. 5–7.

[24] P. Cheng, C.-N. Chuah, X. Liu, Energy-aware node placement in wireless sensor networks, in: Global Telecommunications Conference, GLOBECOM'04, IEEE, 2004, pp. 3210–3214.

[25] Y.T. Hou, S. Yi, H.D. Sherali, S.F. Midkiff, On energy provisioning and relay node placement for wireless sensor networks, IEEE Trans. Wirel. Commun. 4 (2005) 2579–2590.

[26] Y.X.N. Li, S.l. Xie, A power-saving protocol for ad hoc network, in: International Conference on Wireless Communications, Networking and Mobile Computing, 2005.

[27] H. Chih-Shun, T. Yu-Chee, Cluster-based semi-asynchronous power-saving protocols for multi-hop ad hoc networks, presented at the IEEE International Conference on Communications, 2005.

[28] W. Kumar, S. Bhattacharya, B. Qazi, J.M. Elmirghani, A vacation-based performance analysis of an energy-efficient motorway vehicular communication system, IEEE Trans. Veh. Technol. 63 (May 2014) 1827–1842.

[29] S. Guo, O. Yang, Minimum energy multicast routing for wireless ad-hoc networks with adaptive antennas, in: Proceedings of the 12th IEEE International Conference on Network Protocols, 2004, ICNP 2004, IEEE, 2004.

[30] W. Fisher, M. Suchara, J. Rexford, Greening backbone networks: reducing energy consumption by shutting off cables in bundled links, in: Proceedings of the First ACM SIGCOMM Workshop on Green Networking, 2010, pp. 29–34.

[31] M. Gupta, S. Singh, Greening of the Internet, in: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, 2003, pp. 19–26.

[32] G. Fettweis, E. Zimmermann, ICT energy consumption-trends and challenges, in: Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications, 2008, p. 6.

[33] S.-C. Wang, A. Helmy, BEWARE: background traffic-aware rate adaptation for IEEE 802.11, IEEE/ACM Trans. Netw. 19 (2011) 1164–1177.

[34] F.J. Martinez, J.-C. Cano, C.T. Calafate, P. Manzoni, Citymob: a mobility model pattern generator for VANETs, in: IEEE International Conference on Communications Workshops, 2008, ICC Workshops' 08, 2008, pp. 370–374.

[35] I.C. USA, Broadcom Corporation, available: http://www.broadcom.com/products/Wireless-LAN.

[36] P. Viswanath, D. Tse, Fundamentals of Wireless Communication, Cambridge University Press, 2005.

[37] K. TrafficCom, Smarter vehicles, safer roads MCNU R1551, available: http://ww1.prweb.com/prfiles/2008/10/22/915544/MCNUR1551.PDF.

[38] J. Bird, Engineering Mathematics, fifth ed, Elsevier Ltd., 2007.

[39] P. Action, Wind Electricity Generation, online available: http://practicalaction.org/docs/technical_information_service/wind_electricity_generation.pdf.

[40] UK Department for Environment, Food and Rural Affair Dep., UK Air: Air Information Resource [online].

[41] L. Haratcherev, M. Fiorito, C. Balageas, Low-power sleep mode and out-of-band wake-up for indoor access points, in: GLOBECOM Workshops, IEEE, 2009, pp. 1–6.

[42] S.K. Bose, An Introduction to Queuing Systems, Kluwer Academic/Plenum Publishers, 2002.

[43] H. Jung, T.T. Kwon, K. Cho, Y. Choi, REACT: rate adaptation using coherence time in 802.11 WLANs, Comput. Commun. 34 (7/15/2011) 1316–1327.