Energy Aware Scheduling across 'Green' Cloud Data Centres

C. Peoples, G. Parr, S. McClean, P. Morrow and B. Scotney
School of Computing and Information Engineering
University of Ulster
Northern Ireland, United Kingdom
{c.peoples; gp.parr; si.mcclean; pj.morrow; bw.scotney}@ulster.ac.uk

***Abstract*—Data centre energy costs are reduced when virtualisation is used as opposed to physical resource deployment to a volume sufficient to accommodate all application requests. Nonetheless, regardless of the resource provisioning approach, opportunities remain in the way in which they are made available and workload is scheduled. Cost incurred at a server is a function of its hardware characteristics. The objective of our approach is therefore to pack workload into servers, selected as a function of their cost to operate, to achieve (or as close to) the maximum recommended utilisation in a cost-efficient manner, avoiding instances where devices are under-utilised and management cost is incurred inefficiently. This is based on queuing theory principles and the relationship between packet arrival rate, service rate and response time, and recognises a similar exponential relationship between power cost and server utilisation to drive its intelligent selection for improved efficiency. There is a subsequent opportunity to power redundant devices off to exploit power savings through avoiding their management.**

Keywords-Cloud data centre, cost-benefit balance, green policy-based management, operational efficiency, workload scheduling.

I. INTRODUCTION

In 2008, IBM estimated that up to 85% of computing capacity in distributed environments may be idle [1]. It was also estimated that up to 20% of power consumed in data centres can be reduced without impact on runtime as devices, in general, remain powered-on to respond, whether job requests arrive or not [2]. Data centres therefore present an area in which reductions in power drawn have long been achievable. A range of approaches improve data centre efficiency, such as reducing infrastructure, using renewable energy and virtualisation [3]: VMware, Inc., developers of virtualisation technology, estimate up to 80% energy cost reduction through virtualisation [4]. Opportunities remain however, with regard to workload allocation to physical and virtual machines.

Job scheduling is a process ongoing across data centres to accommodate simultaneous reception and response to client requests for throughput maximisation and latency restriction objectives. Scheduling mechanisms have evolved in response to change in the nature and volume of jobs: As the volume of requests has increased and SLA requirements intensify, more sophisticated scheduling mechanisms are based on objectives such as device capabilities on which they run (e.g., I/O scheduling), environment characteristics (e.g., temperature), or stringent application requirements (e.g., activity selection scheduling). Physical hardware characteristics are commonly used to influence workload allocation across clouds and data centres today.

In the process of workload scheduling, data centre management incurs overhead. We define '*efficiency*' in data centres as the occurrence of work performed at servers and satisfaction of client requests when operational costs have been incurred. This acknowledges that costs occur in parallel with server existence in a state where they are able to accept client requests, which may or may not arrive. Costs will also be incurred when supporting operation in terms of the network management system (NMS) and in their physical support through lighting, air conditioning and hardware repair. Inefficient operation therefore occurs where set-up costs accumulated when powering the device on and ongoing management costs are incurred while response to client requests is not. When efficiency is equated with workload processing, server activity should occur to a degree where work performed is of a suitable volume to justify costs incurred.

The focus in this work involves scheduling across data centres such that improved efficiency is achieved. We recognise a unique opportunity in that a unit power cost can drive server (and indeed cloud) selection. This concept is used to optimise performance through 'unload' balancing: In contrast to balancing load across resources, performance benefits through job consolidation on fewer devices and turning off those not required to avoid their management are exploited. This approach is based on the relationship between server utilisation, power cost per job and cost to power and manage a server such that efficiency can be achieved. In using this approach, it is recognised that total server power cost is exponentially distributed with these attributes.

Efficiency and '*justification*' of costs within the scope of this research therefore refers to workload volume processed in relation to total server capacity and its operational cost. In general, costs may be considered to be justified when supporting a high volume of workload, while the opposite is true in instances of low workload. There is a specific utilisation point at which justification occurs, which is most likely to be less than the full rate possible (note that we do not accommodate other factors such as effect of temperature on utilisation and subsequent power inefficiency introduced but acknowledge that other factors influence a cap beyond which utilisation should not increase). It is therefore our objective to define the point at which *justification* occurs and *efficiency* is achieved on a transmission and server-specific basis.

978-3-901882-50-0

The remainder of this paper continues as follows: In Section II the research concept is presented. The effectiveness of the scheme developed is explored in Section III. Finally, the paper concludes and presents future work in Section IV.

II. Research Concept: Efficient load Scheduling

Power is consumed by servers when: 1. Managing the device, 2. Processing application packets, and 3. Operating device hardware. Within the scope of this research, we acknowledge that management costs should be incurred alongside those which achieve the business function, and that those incurred without QoS are inefficient. The objective is therefore to reduce management cost when not responding to application requests. Furthermore, when responding to requests, efficient cost will be achieved at a rate influenced by device physical structure.

A. Problem Definition

Power consumed at server s_m, where m indexes the server, is a function of power cost per port, number of ports a_m, and power consumption per line card l_m (a line card will be applied to one or more server ports), base chassis power b_m, and power cost per byte as a function of port utilisation u_m. This cost is weighted by the power cost associated with the Telecommunications Access Method (TCAM) t_m and port settings p_m [5]. Port settings refer to the line rate forwarding capacity of individual ports. The TCAM refers to packet classification in terms of lookup times. The contribution of ports to server s_m power cost occurs regardless of existence in idle or active states (a port is 'active' when it is involved in data processing, and 'idle' when it is ready to process data but is not actively involved in its processing).

$Power_m$, where m indexes the server, consumed at a device therefore varies as a function of these characteristics:

$$Power_m = \sum_{m=1}^{M} (a_m u_m + l_m + b_m)^{t_m p_m} \quad (1)$$

For the purpose of this research, t_m and p_m are indexed to allow appreciation of the overall server power cost. TCAM and port settings which contribute to higher power cost will be weighted more heavily than those with a lower contribution. The number of ports a_m and line cards l_m, port utilisation u_m and base chassis power b_m are unrelated to TCAM t_m and port settings p_m. The total data centre power cost is a function of the total number of devices M in the data centre, where m is 1 to M servers.

Server utilisation is the probability that its queue is idle or active. More specifically, server utilisation ρ is a measure of the relationship between number of bytes in the queue λ in relation to total byte capacity μ. The queuing system therefore influences the extent to which a server is utilised. When severs in a data centre are in active mode, utilisation ρ may be zero and therefore the power cost for job execution $Power_c$, where c indexes the server, is also zero. (We distinguish this cost from $Power_m$ which is incurred during all server activity and not only when responding to application requests.) Whether ports and line cards are actively processing client requests, they may remain awake to exchange control information. Ports may therefore be awake while utilisation is zero - management activity will occur periodically only. As a further abstraction, utilisation can be distinguished between that achieved when maintaining data centre management (throughput) and when responding to client requests (goodput). Utilisation may therefore occur and costs will be incurred while utility through response to client requests is not. Cost to set up and manage servers, regardless of the degree of utilisation, will be incurred when it exists in any state other than sleep. There is therefore an opportunity to optimise data centre management by improving the probability that utility is achieved when power costs are greater than zero.

The time t_2 to service a packet is a function of queue service rate ∂ in relation to packet arrival rate γ:

$$t_2 = \frac{\partial}{\gamma} \quad (2)$$

Power cost per job $Power_e$ is calculated using the relationship between service rate and job arrival rate, weighted by a power cost Eu charged per unit of the rate serviced:

$$Power_e = Eu(t_2) \quad (3)$$

Server response rate is limited by its total capacity and rate of packet processing through the queue in relation to packet arrival. Power costs incurred before any packet is dropped upon entry into the queue will be wasted costs, incurred without parallel return in achievement of application QoS.

For simplicity, we reduce Equation (1) by considering that power consumed at an active server $Power_b$ is dependent only on cost per job $Power_e$ (which includes the number of active ports and their utilisation) and cost E^D to manage the server such that it is ready to respond to client requests (It is assumed that other elements in Equation (1) remain constant across all devices and need therefore not be considered explicitly.):

$$Power_b = (Power_e + E^D) \quad (4)$$

where $Power_e$ includes cost to respond to an individual job while E^D represents the overall cost to operate and manage the server. This cost is considered because it is assumed, as part of our work, to vary between severs across the data centre based on their physical construction.

We use the approach in [6] to evaluate data centre efficiency using the attributes server set-up and management cost E^D, power cost per job $Power_e$ and server utilisation ρ. The probability distribution function E of power cost $Power_b$ represents the desired server power cost to be achieved overall:

$$\int_0^{Power_b} fE(e)de \geq \delta\% \quad (5)$$

with service within a cost less than $Power_b$ δ % of the time.

Server efficiency in light of its utilisation and cost incurred is measured using the cumulative distribution function (CDF):

$$\int_0^{Power_b} fE(e)de = 1 - e^{-\mu(1-w)Power_b} \quad (6)$$

where w represents the relationship $\frac{Power_e}{\mu}$. Total server power cost is exponentially distributed with the parameters utilisation ρ, power cost per byte s and cost to set-up and manage the server E^D. The CDF describes the probability that the calculation assumes a value less than or equal to utilisation ρ given operational and set-up costs, and reveals the

probability that efficiency, as defined within the scope of this research, will be achieved. A CDF of one reveals that the cost of servicing jobs received will be *justified* and that *efficiency* is met given server utilisation. If a cumulative distribution less than one is achieved, the server is not operating at a utilisation high enough to justify the power cost incurred, and utilisation can be increased to improve efficiency by allocating further jobs within the scope of its residual resources.

Server utilisation is proportional to the rate of port utilisation. Given that we wish to achieve utilisation with an efficient overall power cost, it is possible to determine the number of ports which should support jobs arriving at the server. This investigation therefore also identifies residual port capacity needed to achieve required utilisation and justify operational power cost. Those not required can be turned off so that cost efficiencies from doing so may be accumulated.

The required utilisation will not be achieved immediately upon traffic arrival at the server; there is a delay as utilisation increases due to the finite rate of traffic through ports. Utilising the server to full capacity will therefore take time t_3:

$$t_3 = \frac{ts_c}{tb_c} \tag{7}$$

where ts_c is the total server capacity and tb_c the total number of bits per second when using all server ports v. Once operational, the management system will be responsible for maintaining utilisation at the required rate for efficiency. Server utilisation can be increased by a maximum rate β at any instant, where:

$$\frac{ts_c}{tb_c} = v \tag{8}$$

with tb_c being restricted by the value of v, and:

$$\beta = v(tb_c) \tag{9}$$

with the increase rate limited by full utilisation of all ports. Depending on server utilisation at a point in time and utilisation required to achieve the recommended rate, all ports need therefore not be awake to allow utility and efficiency to be achieved. By intelligently configuring server port state, power cost will be reduced and efficiency achieved according to the contributions of a and u to μ.

B. *Proposed Workload Scheduling Algorithm*

This workload scheduling mechanism is incorporated within a network management system, the energy efficient Context Aware Broker (e-CAB), proposed in earlier work by the authors (e.g., [7]) which encompasses a wider management procedure executed from the point of job arrival at the data centre. When a job arrives at a data centre, the traffic volume is evaluated, server power cost reviewed, current utilisation of servers assessed, and recommended utilisation determined. An assumption made is that data centre resources are partitioned into tiers according to their ability to respond to application QoS [8]. For the purpose of this work, we assume a 3-tier centre: Tier 1 resources accommodate jobs with expedited priority, and tiers 2 and 3 apply a less prioritised service to application requests, varying in the latency with which they respond. There is an order to the way in which jobs are allocated across tiers when operating under constrained conditions (note that close to constrained conditions is our management objective): Providing that there is sufficient capacity at tier 1, a job request typically allocated to a tier 2 or tier 3 server could be allocated to a tier 1 server to exploit the performance improvements from doing so. A job request typically allocated to a tier 1 server on the other hand, is unable to be supported at a tier 2 or tier 3 server where application requirements will be unfulfilled. If there is sufficient availability at a tier 2 server, a job request typically allocated to a t3 server could be serviced here. As part of this design, it is also possible to queue jobs at the management engine: If resources on the required tier are not predicted to become available within a period required by the application, ability to wake sleeping devices is reviewed. When this is not possible, the application will receive, in the worst case, a best-effort service when resources on the optimal tier are not available.

Further decision-making detail is presented in Figure 1, where L is the less expensive server and M the more expensive server. As jobs arrive, the management system allocates them to less expensive servers on the required tier (1). Once the recommended utilisation is reached at this server (2), new jobs may be passed to a more expensive server while capacity in L is released (3) (4). Once utilisation at L is below the recommended level, new jobs may be passed here again (5).

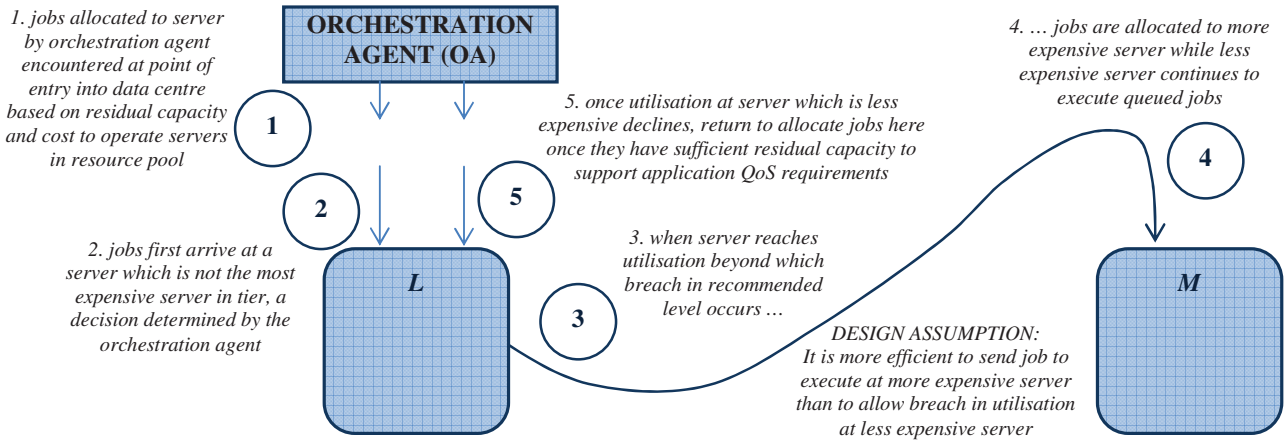

Figure 1 Load Allocation Procedure in Data Centre

This mode of operation is based on the relationship between cost to operate more expensive servers in relation to the cost of breaching recommended utilisation on less expensive servers.

III. Data Centre Optimisation: Case Study

A test scenario demonstrates effectiveness of the approach proposed: A job request has a power cost of 80 Watts and the server has a set-up cost E^D of approximately 150 Watts. Assumptions in this scenario are that jobs are of a constant volume and that their arrival is at a consistent rate across scenarios explored. It is also assumed that server set-up costs and characteristics *l*, *b*, *t* and *p* (see Section II.A) are constant across devices.

The results (Figure 2) demonstrate utilisation required due to costs incurred, validating at a high level that less expensive servers achieve efficiency with lower utilisation given their lower set-up, management and operational costs to compensate for through workload processed. More specifically, Table 1 reveals that server utilisation when there is a power cost per job of 80 Watts and server set-up cost of 150 Watts has to be greater than 81.54% to allow 90% of costs to overcome total expense; server utilisation has to be higher than 85% to achieve efficiency overall. When there is a lower cost of 70 Watts per job on the other hand, utilisation has to be greater than 71.54% to accommodate 90% of power costs; server utilisation has to be higher than 75% to allow operational efficiency for all jobs to be achieved. Once efficiency is achieved, it is possible that utilisation will continue to increase until the maximum limit recommended has been reached – this may be influenced by other characteristics, such as impact of loading on temperature. Performance has also been considered when operating between 10% and 16% utilisation for comparison (Table 2). The CDF disparity from one highlights inefficiency of this utilisation given the power cost distribution.

These results can be used further to extend data centre efficiency through server configuration: Consider, for example, a scenario where the objective is to achieve 85% device utilisation in, for example, a 20 Gigabyte server which uses an adapter with ten Gigabit Ethernet ports *v*. In this example, server utilisation can be increased by a maximum of 62.5% at any point in time, given server and port capacity[1]. If the server, on the other hand, is operating at a utilisation of 50%, it can be increased to operate at the recommended 85% by a 35% volume increase [2], which can be achieved by using six of the ten ports. Four ports may therefore exist in a sleep state, thus improving operational efficiency at this server and avoiding under-utilising a greater number of servers.

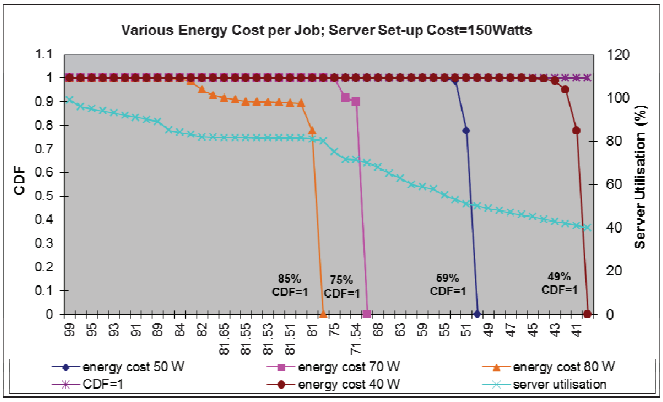

Figure 2 Various Cost per Job; Server Set-up Cost 150W

Table 1 CDF: Job Power Cost 80 Watts, Server Set-up Cost 150 Watts (Utilisation 80-95%)

Utilisation	80	81	81.5	81.55	82	85	89
CDF	0	0.78	0.89	0.90	0.95	0.99	1

Table 2 CDF: Job Power Cost 80 Watts, Server Set-up Cost 150 Watts (Utilisation 10-16%)

Utilisation	10	11	12	13	14	15	16
CDF	-3.98	-8.90	-1.98	-4.43	-9.89	-2.2	-4.92

IV. Conclusion & Future Work

In integrating network management, additional power cost is incurred. In this paper, we present a mechanism to achieve green network management and propose that, when allocating jobs to servers, decisions use the CDF to achieve a server utilisation which justifies operational cost. Furthermore, results demonstrate efficiency optimisation by controlling server configuration in response to the required utilisation which accommodates the recommended operation volume.

While optimised operation in data centres is demonstrated by exploring relationships between workload and efficiency, other attributes can also be used to perhaps improve selection accuracy, such as, for example, server RPM. Furthermore, application QoS characteristics could also be included, such as the optimal response latency or acceptable packet loss rate. These ideas will be explored as part of future work.

ACKNOWLEDGEMENT

This research is supported by the IU-ATC (www.iu-atc.com), funded by the UK EPSRC Digital Economy Programme and Government of India Department of Science and Technology.

REFERENCES

[1] IBM Corporation, "IBM's Strategy for Dynamic Infrastructure," 2008.

[2] Intel, "Data Center Energy Efficiency with Intel Power Management Technologies," Intel Information Technology, Feb 2010.

[3] BT, "A Realist's Guide to Green Data Centers," 2008.

[4] VMware, Inc., "How VMware Virtualisation Right-sizes IT Infrastructure to Reduce Power Consumption," White Paper, 2011.

[5] V. Manral, "Benchmarking Power Usage of Networking Devices," IETF 'work in progress' as an Internet Draft, Jan. 2011.

[6] K. Xiong and H. Perros, "Service Performance and Analysis in Cloud Computing," in Proc. World Conf. on Services, Jul. 2009, pp. 693-700.

[7] C. Peoples, G. Parr and S.McClean, "Context-aware Characterisation of Energy Consumption in Data Centres," in Proc. 3rd IFIP/IEEE Int. W'shop on M'ment of the Future Internet, May 2011, pp. 1250-57.

[8] M. Karir, "Data Centre Reference Architectures," IETF 'work in progress' as an Internet Draft, Oct. 2011.

[1] Maximum potential increase in utilisation given server and port capacity:

$$\frac{(2E+10*8)/100}{10*1E+10*10} = 62.5 \quad (10)$$

[2] Number of ports which require utilisation to increase server capacity to recommended rate:

$$\frac{((2E+10*8)/100)*35}{10*1E+10} = 5.6 \quad (11)$$