International Journal on Cloud Computing: Services and Architecture (IJCCSA),Vol.2, No.1,February 2012

A perspective approach on Energy efficiency and Cloud Faster of mobile in cloud computing

Vijeyta Devi, Vikas Kumar, Manjula Poojary, Dr. Vadlamani Nagalakshmi and Narasimham Challa

Dept.of Computer Science & Engineering
VITAM College of Engineering & Technology, Visakhapatnam, A.P

`ram_yadav536@yahoo.com`

Abstract

In this paper we propose an analysis of the critical factors affecting the energy consumption of mobile clients in cloud computing and consideration to make performance very fast by using Cloud faster tool to solve speed problems. This can be done by increasing throughput, minimizing response time, and increasing resource utilization. Energy middle ware solutions are used to evaluate the feasibility of automatic decision making between local and remote processing. Cloud two suits are DCTCP and Wide Area TCP. If we compare DCTCP delivers better throughput than TCP, while using 90% less buffer space but TCP provides high burst tolerance and low latency for short flows.

Keywords

Cloud computing, Energy trade-off analysis, Communications in Commoditized data centers, Cloud faster, TCP.

I. Introduction

In cloud computing energy efficiency have become challenges to make it capable technology, which can offer numerous benefits for battery powered mobile devices by saving energy of batteries. At present the mobile characteristic will decide the future of cloud technology and for mobile devices energy management is equally important for their prolong life. So, energy efficiency of mobile devices have become equally important as present cloud computing does not allow running existing applications with less energy. In this paper we have discussed the computing to communication ratio, which is the critical factor for the decision between local processing and computation offloading. With help of preliminary results on mechanisms for calculating the cost of modern web oriented workloads and data centers, we have found that these applications are realistic and but requires require three things from the data center network which are low latency for short flows, high burst tolerance, and high utilization for long flows. After that we will see how things can be achieved. First two of them requires Partition/Aggregate workflow pattern and high throughput for long flows that update the data is the third essential thing for low latency and burst tolerance. DCTCP is effective in achieving full throughput by taking up a very small footprint in the switch packet buffer, as compared to TCP. There are many differences between the data center environment and wide area networks (WANs), where most of the prior work on TCP has focused. TCP is very vast and have two families of congestion control protocols that attempt to control queue lengths which are Implicit delay-based protocols use in RTT measurements and another one is Active Queue Management (AQM) approaches which uses explicit feedback from congested switches. This helps us in understanding that DCTCP provides all the benefits we seek.

DOI : 10.5121/ijccsa.2012.2103

2. Cloud Computing

Recently Cloud computing has received large interest. The primary motivations for the mainstream cloud computing are related to the elasticity of computing resources. This offers us virtually infinite resources that are readily available on demand and charged according to usage, which has become economic advantages both for cloud providers and cloud users. The motivations intended for mobile devices users for cloud computing differ from the motivation of cloud computing with well-connected PC devices. The main bottleneck between mobile and mainstream cloud computing is the data transfer. Here we are using small devices which will get lost easily but it requires transferring large amount so device backup is one of useful service. So, synchronization of contact and calendar data, where the amount of transferred data is more modest, is a service that is widely available. Now for all these mobile devices energy efficiency is another big challenge. For that use cases have been developed towards always on-line connectivity, high speed wireless communication, high definition multimedia, and rich user interaction. For these devices battery technology has not been able to match the power requirements for the increasing resource demand. This can be attributed to amount of energy that can be stored in a battery, which is limited and is growing only 5% annually. If we start using bigger batteries than it will result into larger devices that is not an attractive option. For better energy efficiency we have to work on other benefits like device size, cost and R&D efficiency. This can be easily seen from the fact that, large parts of the hardware technology benefits that have been traded for programmability in mobile phone designs. With the use of modeling and simulation studies, an investigation of offloading work from mobile to a fixed host has carried out and found that under certain conditions 20% energy savings would be possible. Compiler technology has been studied in, e.g. where a program is partitioned to client and server parts. The main metrics evaluated are execution speed and energy consumption. There are numerous possibilities by which significant energy savings are possible. But executions of all these requires accounting of resources , execution time , energy usage, criteria for deciding local, remote and hybrid execution by virtual machine technology for mobile cloud technology. In this paper we propose distributed cloud architecture for utilizing single hop radio technology for overcome the high latency and jitter. But implementation of proposed architecture requires huge changes in infrastructure. We also try to investigate both wireless local area network and cellular communication to find out the thresholds for moving towards cloud technology.

3. Analysis for energy efficiency.

In cloud computing, the critical aspect for mobile clients is the trade-off between energy consumed by computation and the energy consumed for communication. We need to consider the energy cost of performing the computation locally (Elocal) versus the cost of transferring the computation input and output data (Ecloud). For offloading to be beneficial we require that

$$\text{Ecloud} < \text{Elocal} \qquad (1)$$

If D is the amount of data to be transferred in bytes and C is the computational requirement for the workload in CPU cycles then

$$\text{Ecloud} = \frac{D}{\text{Deff}} \qquad (2)$$

$$\text{Ecloud} = \frac{C}{\text{Ceff}} \qquad (3)$$

where Deff and Ceff are device specific data transfer and computing efficiencies. The Deff parameter is a measure for the amount of data that can be transferred with given energy (in bytes per joule) whereas the Ceff parameter is a measure for the amount of computation that can be performed with given energy (in cycles per joule). With these we can derive the relationship between computing and communication for offloading to be beneficial

$$\frac{C}{D} > \frac{Ceff}{Deff} \tag{4}$$

Device/frequency	**Power/W**	**Cycles/energy (Ceff)**
N700/400 MHz N700/330 MHz N700/266 MHz N700/165 MHz	0.8 0.7 0.5 0.3	480 MC/J 480 MC/J 540 MC/J 510 MC/J
N800/600 MHz N800/550 MHz N800/500 MHz N800/250 MHz	0.9 0.8 0.7 0.4	650 MC/J 690 MC/J 730 MC/J 700 MC/J

Table 1: Energy characteristics of local computing for Nokia N700 and N800 (MC=megacycle).

The computing energy efficiency (Ceff) is affected by the device implementation. For example, a CPU designed for high peak performance requires much more power per megahertz than a core designed for lower performance. Techniques like dynamic voltage and frequency scaling (DVFS) alter the power and performance of the CPU at run-time. Table 1 lists the computational energy characteristics of two mobile devices, the Nokia N700 and Nokia N800, measured with the gzip deflate compression program compressing ASCII data. In this paper we have used the Nokia Energy Profiler, which measures the power consumption of the complete device. Outcome of this study is that the power and bit-rate characteristics of wireless modems depends upon the activity time of interface. The latencies associated with the activation and deactivations of the wireless interface vary by technology and are longer in cellular communication than in WLAN.

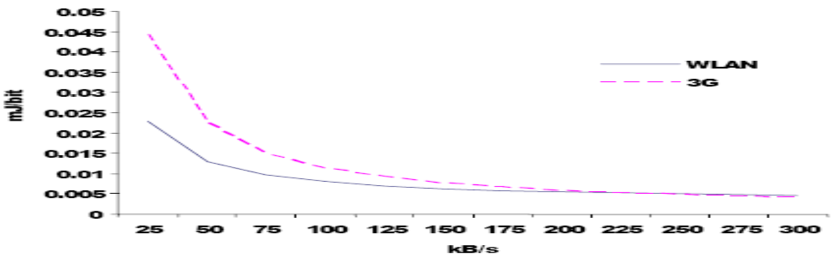
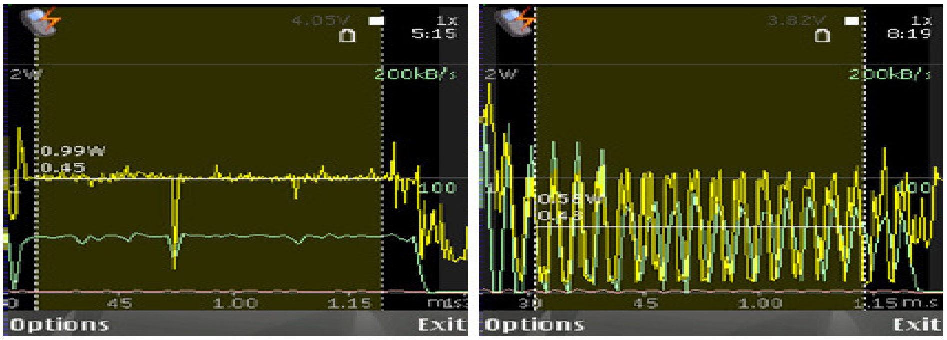


Figure 1: Energy per bit for N90 WLAN and 3G.

In Figure 1 the dependency of energy per transferred data on communication bit-rate is illustrated. If the bit-rate is higher than the data transfer will be more energy efficient. In this figure it is also illustrated that the energy efficiency of cellular communication tends to be more sensitive to the data transfer bit-rate than WLAN



(a) Smooth traffic source. (b) Bursty traffic source

Figure 2: Traffic pattern effect for N90 WLAN.

In Figure 2, it is shown that the traffic pattern affects the energy efficiency of communication and also shows that data transfer power levels for WLAN communication with smooth and bursty traffic sources measured on Nokia N90. It is apparent from above figure that the average bit-rate that requires 1W power with a smooth traffic source consumes only 0.6W with a bursty traffic source.

Device	**Power/W**	**Bytes/energy (Deff)**
N700/400 MHz N700/330 MHz N700/266 MHz N700/165 MHz	1.5 1.4 1.3 1.1	390 KB/J 370 KB/J 350 KB/J 310 KB/J
N800/WLAN N800/3G/receive (near) N800/3G/transmit (near) N800/3G/receive (far) N800/3G/transmit (far)	1.1 1.1 1.0 1.4 3.2	860 KB/J 450 KB/J 190 KB/J 350 KB/J 60 KB/J

Table 2: Nokia N700 and N800 wireless transfer Energy characteristics.

In Table 2 record of the energy characteristics of wireless communication for the Nokia N700 and N800 with the netperf TCP streaming benchmark. The CPU operating point affects significantly the WLAN throughput of the N700.that is why the metrics are shown for all operating points separately. Even though the power level of the device is highest at the highest operating point, higher throughput causes it to be the most energy efficient state for WLAN transfers.

Workload	Cycles/byte
gzip ASCII compress x264 VBR encode	330
x264 CBR encode html2text	1300
wikipedia.org html2text	1900
en.wikipedia.org pdf2text	2100
N800 data sheet pdf2text	5900
E72 data sheet	960
	8900

Table 3: Data ratios for various workloads Computation.

The CPU operating point will not affect the N800 networking throughput. In this table the effect energy efficiency of the data transfer from near and far from base station for N800 is shown. Here in this table the fact that the 4G measurements are done in a crowded network are done where the devices will not be able achieve optimal throughput and also kept in consideration for the asymmetric design of the 4G HSDPA communication. Here it also illustrated that the power consumption of the wireless modem is significantly affected by the network quality. The cost of transmitting is much higher than receiving. When transmission far from the base station, which requires very high power degrading the energy efficiency of the communication. Combining the best case values of Tables 1 and 2 with Equation 4 we get information for computation offloading which is beneficial for the workload needs to perform more than 1000 cycles of computation for each byte of data. The effect of processed data is especially significant for modern web oriented workloads where the content largely dictates the processing requirement.

Device/bearer	Average power/W	Total energy/J
N700/local N700/WLAN N800/local	1.2	50
N800/WLAN N800/3G (near)	1.4	20
N800/3G (far)	1.4	40
	1.2	20
	1.5	40
	2	55

Table 4: Power and energy of PDF viewing.

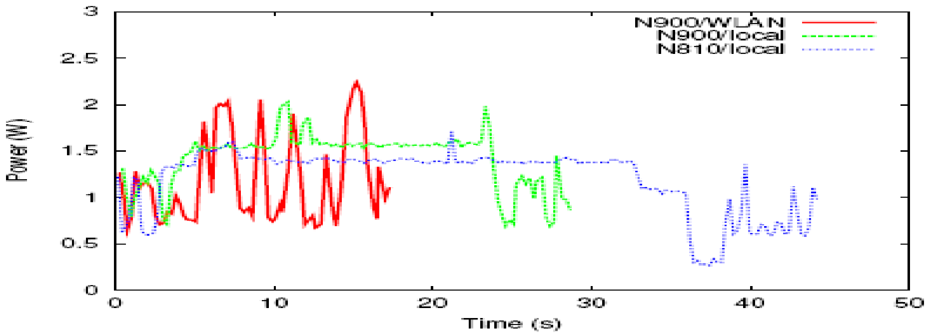


Figure 3: Example power curves for PDF viewing.

3.1 Example: Mobile as a Thin Client

We have made experiments with PDF viewing and web browsing where the mobile terminal acts as a thin client utilizing the window system. The network transparency feature allows running the application and its display in separate devices. This mechanism is available in the maemo platform used by the Nokia N700 and N800 devices. Table 4 shows the average power levels and total energies for viewing a demanding PDF document (Nokia E72 data sheet). For N700 there are two cases: local viewer and remote X11 client connected over WLAN. For N800 there are four cases: local viewer, remote viewer over WLAN, remote viewer over 4G packet data near base station and remote viewer over 4G packet data far from base station. Figure 4 shows as examples the measured power for local viewer cases and the N800 WLAN case. As can be seen, the remote cases run with higher average power. However, the total energy for the remote WLAN case is the smallest because of shorter execution time. The 3G network cases consume more energy than WLAN because of communication latencies. Moreover, 4G Communication is sensitive to location as evidenced by the energy differences of the cases where the mobile device is near and far from the base station.

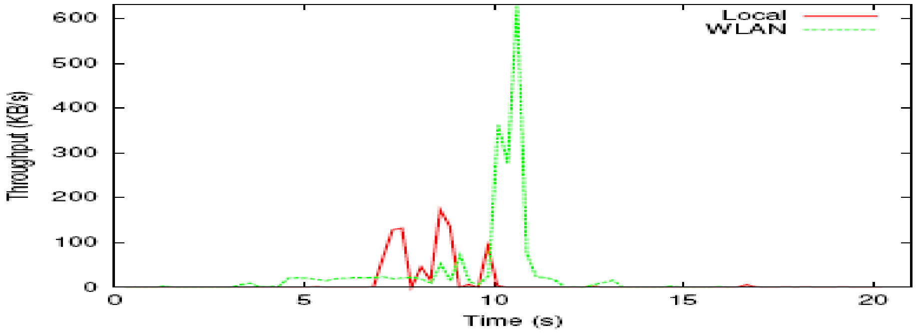


Figure 4: Download rate for local and remote browser

Another observation that can be made from the measurements is the improved processing performance of the N800 compared to the N700. The Nokia N800 is a more recent model than N700 and has significantly more powerful processing capabilities. Even though the N800 local processing takes a bit higher power, the shorter processing time makes the N800 more energy efficient for this workload. For the WLAN case both devices perform similarly regardless of the better WLAN throughput that N800 is able to achieve. In the web browsing case, we compared local browser with remote X11 client browser. Even though web page processing requires significant computation, and is therefore a good candidate for offloading, the remote web browser causes also significant amounts of network traffic while rendering the page. Figure 3 shows example curves of the download traffic for local and remote browser. Even though the local browser downloads less data (about 200KB) than the remote case (about 500KB), the remote browser case achieves higher throughput for a large part of the data transfer. This highlights the fact that the amount of transferred data alone is not a sufficient metric for characterizing the energy consumption of communication. For X11 applications, the usage pattern and application implementation details have also major impact on the energy consumption. Rendering requires significant computation and running it remotely can therefore save energy. However, for example scrolling requires minimal processing and can cause large amounts of data to be transfer which causes remote operation to be very energy inefficient in addition to degrading interactive performance.

4. Data Centers communication

In this context first we will understand the challenges for data center transport protocols. After that why latency is a critical metric in data centers. All web search, social network content composition, and advertisement selection are based on design pattern applications. To make interactive, soft-real-time applications, latency is the key metric. Customer impact studies decide the total permissible latency. An all-up SLA is calculated after subtracting typical Internet and rendering delays, the “backend” part of the application is typically allocated between 230-300ms.

4.1. TCP & DCTCP Workload Characterization

We will now see the basic understanding why TCP behaves poorly and for the creation of benchmarks for evaluating DCTCP. With total of over 6000 servers in over 150 racks the three clusters support soft real-time query traffic, integrated with urgent short message traffic that coordinates the activities in the cluster and continuous background traffic that ingests and organizes the massive data needed to sustain the quality of the query responses.

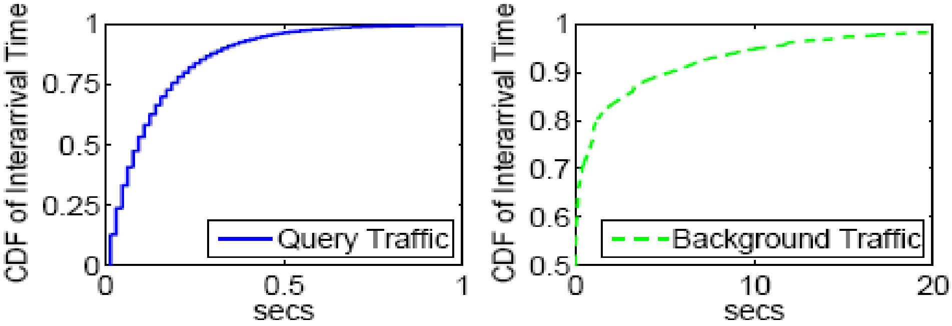


Figure 5: CDF of Time Vs arrivals of new work for the Aggregator (queries) and between background flows between servers (update and short message).

Figure 5 shows the time between arrivals of new background flows. The inter arrival time between background flows reflects the superposition and diversity of the many different services supporting the application. The query traffic consists of very short, latency-critical flows, following a relatively simple pattern, with a high-level aggregator partitioning queries to a large number of mid-level aggregators.

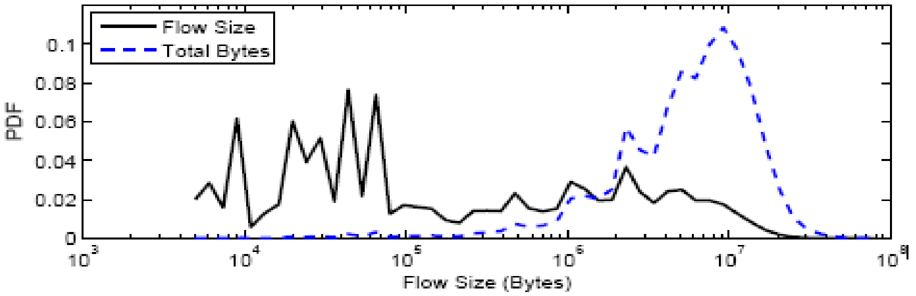


Figure 6: PDF Vs Flow size distribution for background traffic.

4.2 Background Traffic

Query traffic is a complex mix of background traffic, consisting of both large and small flows. Figure 6 presents the PDF of background flow size, illustrating how most background flows are small, but most of the bytes in background traffic are part of large flows

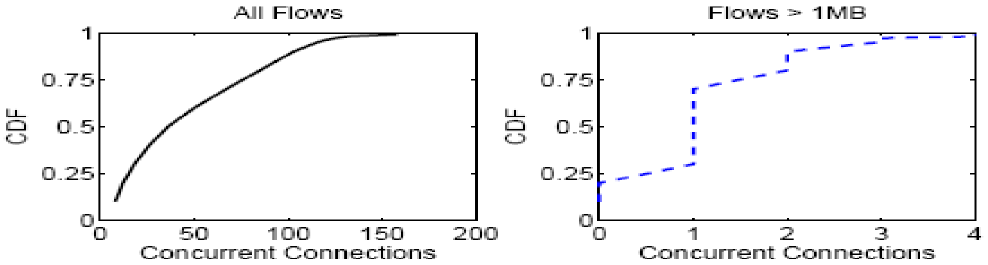


Figure 7: Distribution of number of concurrent connections.

4.3 Flow Concurrency and Size

Figure 7 presents the CDF of the number of flows vs distribution of concurrent connections. The 99.99th percentile is over 1,600, and there is one server with a median of 1,200 connections. When only large flows (> 1MB) are considered, the degree of statistical multiplexing is very low the median number of concurrent large flows is 1, and the 75th percentile is 2. Yet, these flows are large enough that they last several RTTs, and can consume significant buffer space by causing queue buildup. In summary, throughput-sensitive large flows, delay sensitive short flows and bursty query traffic, co-exist in a data center network. Yet, these flows are large enough that they last several RTTs, and can consume significant buffer space by causing queue buildup. In summary, throughput-sensitive large flows, delay sensitive short flows and bursty query traffic, co-exist in a data center network.

5. Architecture and Protocols for Cloud Mobile Faster

To make cloud mobiles fast we have developed a new suite of architectures and protocols that boost performance and the robustness of communications to overcome speed problems. The results are backed by real measurements and a new theory describing protocol dynamics that enables us to repair fundamental problems in the Transmission Control Protocol. We have developed two suites of technology DCTCP and Wide Area TCP. In the congestion control algorithm of TCP changes has been done to decreases application latency inside the data center by decreasing queue lengths and packet loss while maintaining high throughput. By doing changes to the network stack of the "last-hop server" - the last server to touch packets before they travel to the client - that reduce the latency for transferring small objects by working around last-mile impairments such as loss and high RTT.

6. DCTCP Algorithm

It is designed to operate with very small queue occupancies and without loss of throughput. By using DCTCP we can achieve high burst tolerance, low latency, and high throughput, with commodity shallow buffered switches. DCTCP overcomes congestion by reacting to congestion in proportion to the extent of congestion by delay-based congestion algorithms. DCTCP uses a very simple marking scheme at switches that sets the Congestion Experienced (CE) code point of packets as soon as the buffer occupancy exceeds a fixed small threshold. The DCTCP source reacts by reducing the window by a factor that depends on the fraction of marked packets: the larger the fraction, the bigger the decrease factor.

6.1 Algorithm

The DCTCP algorithm has three main components

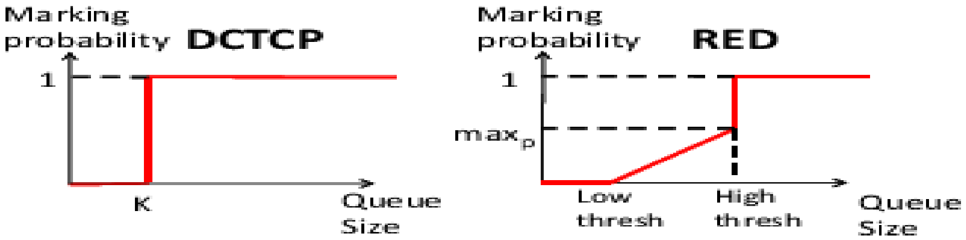


Figure 8: DCTCP's AQM scheme is a variant of RED: Low and High marking thresholds are equal, and marking is based on the instantaneous queue length.

6.1.1 Marking at the Switch

It employs a very simple active queue management scheme, shown in Figure 8. There is only a single parameter, the marking threshold, K. An arriving packet is marked with the CE code point if the queue occupancy is greater than K upon its arrival. Otherwise, it is not marked. This design marking scheme is motivated by the need to minimize queue buildup.

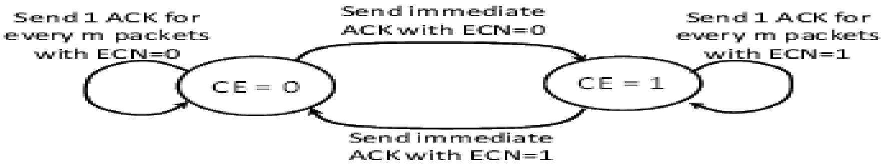


Figure 9: Stages for ACK generation state machine.

6.1.2 ECN-Echo at the Receiver

The main difference between a DCTCP receiver and a TCP receiver is that the CE code points are conveyed back to transmitter in different ways. RFC 3168 states that a receiver sets the ECN-Echo flag in a series of ACK packets until it receives confirmation from the sender that the congestion notification has been received. Also, the DCTCP receiver send the same packets back to sender.The simplest way to do this is to ACK every packet, setting the ECN-Echo flag if and only if the packet has a marked CE code point.

6.2 Benefits

6.2.1 Congestion

By using DCTCP we can achieve high burst tolerance, low latency, and high throughput, with commodity shallow buffered switches. It overcomes congestion by reacting to congestion in proportion to the extent of congestion by delay-based congestion algorithms

6.2.2 Incast

The incast scenario, where a large number of synchronized small flows hit the same queue, is the most difficult to handle. Because DCTCP starts marking early (and aggressively – based on instantaneous queue length), DCTCP sources receive enough marks during the first one or two RTTs to tame the size of follow up bursts, preventing buffer overflows.

6.2.3 Utilizing single hop radio technology

A distributed cloud architecture utilizing single hop radio technology for reducing latency and jitter. However, the proposed architecture requires significant changes infrastructure existing

6.2.4 Resources are available on demand and charged according to usage

All infrastructure required for cloud mobile computing is easily available and can be charged according to usage. This offers considerable economic advantages both for cloud providers and cloud users .Cloud computing has the potential to save mobile client energy but the savings from offloading the computation need to exceed the energy cost of the additional communication.

7. DISCUSSION

The setup for mobile cloud computing is substantially different from the traditional client-server computing arrangement. Energy is a fundamental factor for battery powered devices and an important criterion when considering moving computing to the cloud. The basic balance between local and remote computing is defined by the trade-off between communication energy and computing energy. However, there are many factors to consider when thinking about mobile cloud computing scenarios. The computing to data ratio defining the break-even for moving to cloud is highly dependent on the exact energy efficiencies of wireless communication and local computing. The measurements in this paper provide a rough guideline for current mobile devices but technology development can shift the trade-off point significantly.Naturally device specific implementation decisions affect the balance but to a less radical extent. Also, the computation offloading needs careful design in order to avoid introducing long latencies into user visible operations. As shown in Figure 1, computation offloading can in some cases be used to improve performance in addition to saving energy. For wireless communication, not only the amount of data but also the communication pattern has a large impact on the energy consumption. E.g., interactive workloads utilizing thin client technologies represent probably the most challenging target because of the fine granularity of the required communication. The best energy efficiency for communication would be achieved with bulk data transfers. Also if immediate response is not required the data transfers can be delayed and executed later when a bearer with better energy efficiency is available. Scheduling data transfers to happen in parallel can also be used to save energy Mobile cloud computing differs also from simple computation offloading in the sense that the cloud can offer services other than computing for mobile clients. Storage for backing up the mobile terminal data is one example. Another example could be a content sharing service, which by nature requires transferring locally produced data to the cloud. Also, for many use cases the data is already in the cloud (e.g., web content). In this kind of scenarios the cost of transferring workload input is essentially zero. However, signaling traffic for controlling the computation would still be needed. It is clear that there are a number of nontrivial factors to consider when making design decisions about cloud applications targeting mobile devices.

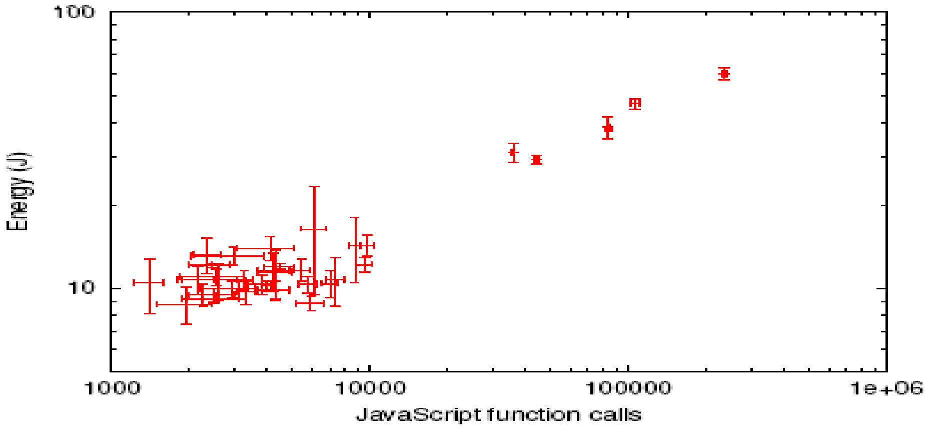


Figure 311: Energy versus JavaScript function calls (in logarithmic scale on both axis to accommodate loads of different magnitude).

Estimating the computational requirements of client side processing and the energy consumption of the required network traffic is therefore an important topic. Currently web technologies are a popular way of constructing distributed applications and web applications are increasingly targeting mobile clients. There is a need for energy consumption feedback during the natural development and debugging cycle but current tools are severely lacking in this area. very simple metrics allow coarse grain estimation of the energy consumption of JavaScript execution. DCTCP, we evaluated Active Queue Management (AQM) schemes like RED and that do not modify TCP's congestion control mechanism. We found they do not work well when there is low statistical multiplexing and traffic is bursty—both of which are true in the data center.

8. CONCLUSIONS

In this paper we have analyzed the energy consumption of mobile clients in cloud computing. This is especially true for wireless communication where achieving high energy efficiency requires high throughput. It is also important to realize that the performance metrics of real world scenarios can be significantly different from theoretical maximums implied by device components. Context dependency of the energy efficiency trade-offs means that the decision making cannot be restricted to design time only. Energy aware middleware solutions should therefore be researched to evaluate the feasibility of automatic decision making between local and remote processing. For interactive workloads, latencies associated with wireless communication are a critical factor. Implementing modern rich user interfaces for cloud applications with high energy efficiency is an especially challenging topic.The energy consumption of a mobile device is affected by the complete end-to-end chain. Optimizing wireless communication patterns is critical for energy efficiency and requires considerations both on the client and the server side. Found that to meet the needs of the observed diverse mix of short and long flows, switch buffer occupancies need to be persistently low, while maintaining high throughput for the long flows.

ACKNOWLEDGMENTS

For support and review work we thank our team at Vitam college of Engineering & Ph.d candidates of C.S.E Dept Science and Technology. For our research tools we are grateful to numerous people my known in Nokia product development units & IBM.

REFERENCES

[1] NEUVO, Y. Cellular phones as embedded systems. Digest of Technical Papers, IEEE Solid- State Circuits Conference (ISSCC) (2004), 32–37

[2] A. Aggarwal, S. Savage, and T. Anderson. Understanding the performance of TCP pacing. In IEEE Infoc.om2000, 2001

[3] P.M. Chen and B.D. Noble. When virtual is better than real. Proceedings of the 2009 IEEE on Hot Topics in Operating Systems (Hot OS), 26(5) (1998) 242–261

[4] L.P. Cox and P.M. Chen. Pocket Hypervisors: opportunities and Challenges. IEEE Transactions Proceedings of Hot Mobile.28 (5) (1998) 282-340

[5] F-Secure Corporation. F-secure mobile anti-virus. http://mobile.f-secure.com

[6] Jason Flinn, Dushyanth Narayanan, and M .Satyanarayanan .IEEE Self-tuned remote execution for pervasive computing. In Proceedings

[7] Fuggetta, A., Picco, G.P., Vigna, G.: Understanding code mobility. IEEE Transactions on Software Engineering 24(5)(1998) 342–361

[8] L.P. Cox and P.M. Chen. Pocket Hypervisors: Opportunities and Challenges. IEEE Transactions Proceedings of Hot Mobile. 28(5) (1998) 282-340.

[9] F-Secure Corporation. F-secure mobile anti-virus. http://mobile.f-secure.com.

[10] Jason Flinn, Dushyanth Narayanan, and M. Satyanarayanan. IEEE Self-tuned remote execution for pervasive computing.

[11] Fuggetta, A., Picco, G.P., Vigna, G.: Understanding code mobility. IEEE Transactions on Software Engineering 24(5) (1998) 342–361.

[12] A. Aggarwal, S. Savage, and T. Anderson. Understanding the performance of TCP pacing. In IEEE Infocom 2000, 2000

Authors

Vijeyta Devi: Working as a Associate Professor in CSE Department at VITAM College of Engineering and Technology, Sontyam, Anadapuram Mandal, Visakhapatnam, Andhra Pradesh, India. She has 5 year Teaching Experience and published more than 5 Papers in reputed National International, IEEE and Journal

Vikas Kumar: He is Operational Researcher and published more than 4 papers in International Journal and IEEE.

DR. Vadlamani Nagalakshmi: working as a professor in GITAM University, Visakhapatnam, Andhra Pradesh, India. She has 11 year Teaching Experience and 6 year Research Experience. She published more than 19 Papers in reputed National, International, IEEE and Journal