Deliverable DA2.1.15
Conference paper on holistic device-level energy and power management

Jukka Nurminen, Jussi Ruutu, and Sasu Tarkoma

ICT SHOK Future Internet Programme
(ICT SHOK FI)

Phase 2: 1.6.2009 – 31.12.2010

Tivit, Yritysten tutkimus- ja kehittämisrahoitus, Päätös 516/09, 29.5.2009, Dnro 560/31/09

TKK, Tutkimusrahoituspäätös 40212/09, 29.5.2009, Dnro 925/31/09

www.futureinternet.fi

www.tivit.fi

This work was supported by TEKES as part of the Future Internet programme of TIVIT (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT).

The Case for Holistic Mobile Energy and Power Management

Jukka K. Nurminen
Nokia Research Center
Helsinki, Finland
jukka.k.nurminen
@nokia.com

Jussi Ruutu
Nokia Research Center
Helsinki, Finland
jussi.ruutu@nokia.com

Sasu Tarkoma
Nokia Research Center and
University of Helsinki
Helsinki, Finland
sasu.tarkoma@nokia.com

ABSTRACT
In this paper we consider the dimensions of mobile Energy and Power Management (EPM), outline an EPM design, and elaborate holistic EPM opportunities for battery-operated mobile devices. We argue that a holistic EPM design can offer significant increases for battery lifetime. Our argumentation is based on a number of cases that illustrate how energy consumption can be reduced by opportunistically utilizing available resources, either in the distributed system or in the local operating context. The proposed holistic EPM design is based on multi-dimensional optimization over three different EPM dimensions (vertical, horizontal, and temporal) and their respective EPM mechanisms. We illustrate the holistic EPM concept and the collaboration of different EPM dimensions via a content downloading example.

General Terms
power management, energy management, mobile computing

1. INTRODUCTION
The energy consumption is one of the greatest challenges for current mobile devices. Energy and power continue to remain the most limiting factors for the performance of mobile computing systems. The battery capacity increases approximately 10% annually while the requirements increase in a more rapid pace. In several consumer surveys the duration of the battery charge has been named as one of the top five problems. The near future does not look any better; the increasing complexity of computing platforms and new services increase the energy consumption requirements. Recent developments pertaining to energy source have not kept pace with the increasing requirements. Especially Internet and Web 2.0 service usage consume vast amounts of energy and result in short battery lifetimes, and ultimately poor user experiences.

Energy is a significant limiting factor in determining the performance of mobile devices. There is significant need for reducing energy consumption, which can be traded for more operating time, better performance, or smaller size of the device. The modern mobile device operating environment is complex and typically involves a significant number of hardware and software components, some of which can be profiled and instrumented from the power management point of view. For example, it might be possible to change from one wireless link to another, or to utilize them at the same time in order to complete a data transfer as soon as possible and then set the wireless interfaces to lower-power states. Moreover, it might be possible to offload computation to a service on the Internet (the Cloud) in order to conserve battery life. These examples give an idea of the possibilities and trade offs involved with Energy and Power Management (EPM).

In this paper we consider the dimensions of mobile energy and power management, outline an EPM design, and elaborate several holistic EPM opportunities. We argue that a holistic EPM design can offer significant increases for battery lifetime. Our argumentation is based on a number of cases that illustrate how energy consumption can be reduced by opportunistically utilizing available resources, either in the distributed system or in the local operating context. The proposed holistic EPM design combines three different EPM dimensions (vertical, horizontal, and temporal) and their respective EPM mechanisms. EPM decisions may be made at the design or run time. At the moment the utilization of multiprocessing, cloud computing, smart spaces and application-level collaboration in resource usage (e.g. data communication, CPU) are the most prominent opportunities.

2. DIMENSIONS OF EPM
As the importance of energy-efficiency of wireless devices is widely recognized, significant amount of work has been done on this area. Work has been done at different levels ranging from component technology to applications, and a number of tools have been developed for energy consumption monitoring and analysis. Currency, accounting and other dynamic economic models with energy focus have been proposed and integrated with OS design. For instance, the Stanford's Cinder is an example of a recent OS proposal that integrates energy accounting and limiting [6]. However, most of the EPM work is highly focused; the target is to improve the energy-efficiency of processing, radio communication, display usage, or some other task.

In our view, all of these different research perspectives are important; however, significant gains are expected to be achieved when the different approaches are not developed and used in isolation but in a combination. In essence a co-operative EPM framework is needed that works across multiple levels, including hardware, software, and user experience.

As evidenced by the opportunities that cross-layer design has created in communications, we believe that wider consideration of different perspectives should also be useful for the energy-efficiency of handheld devices. We divide the EPM into three dimensions: horizontal, vertical, and temporal. Figure 1 shows our model of the different dimensions of EPM. As we will discuss, the opportunities of EPM are not limited to the individual dimensions but to the appropriate combination of the different axis.

2.1 Vertical Dimension

This is the traditional dimension for EPM in embedded devices. As presented in Figure 1, the device has several vertical layers (from transistor to user) each of which tries to optimize their own energy consumption in some way. Some prior art exists to increase interoperability between vertical layers (e.g. Advanced Configuration and Power Interface [1]), but in practice the EPM algorithms and mechanisms often consist of a set of local optimizations at different layers without proper system level coordination. The policies targeting local optima in each layer are not likely to lead to the global minimum for energy consumption. For example, too active sleep mode at wireless interface layer may cause retransmissions at Internet network protocol layer due to unpredictable latency. This in turn, may cause longer active time for CPU and memory and actually increase the energy consumption.

2.2 Horizontal Dimension

The horizontal dimension offers possibilities for EPM in mobile systems by using parallelism at various levels. Multicore processors provide parallelism at CPU level, while off-loading computing to other devices near-by or in the cloud provide parallelism on a higher level. Parallel operations within the device typically decrease the power consumption, because the tasks will be finished earlier and the components can go to sleep, or the tasks can be accomplished in the same time with lower power consumption (e.g. running processing elements with lower speed and power).

Distributing, off-loading and sharing computing will move some of the computation away from the device by removing the need to spend power computing the off-loaded tasks completely. However, it is important to note that in off-loading we are typically trading computing costs with communication cost; transferring data can require more energy that the computation that can be avoided by off-loading. It should also be considered that off-loading can cause decreased user experience, if designed poorly. Thus, careful optimization is needed. However, distributing computation may become increasingly popular due to the fact that the content also becomes more distributed.

2.3 Temporal Dimension

In the temporal dimension the aim is to perform activities at the right time. Scheduling is often investigated from the execution speed and real-time point of view, but it can also be used to optimize energy consumption. The basic mechanism for temporal axis is to schedule tasks to maximize the sleep periods in low-power states. This can be achieved by adjusting timers appropriately, by grouping a number of activities together, or by simply eliminating some tasks.

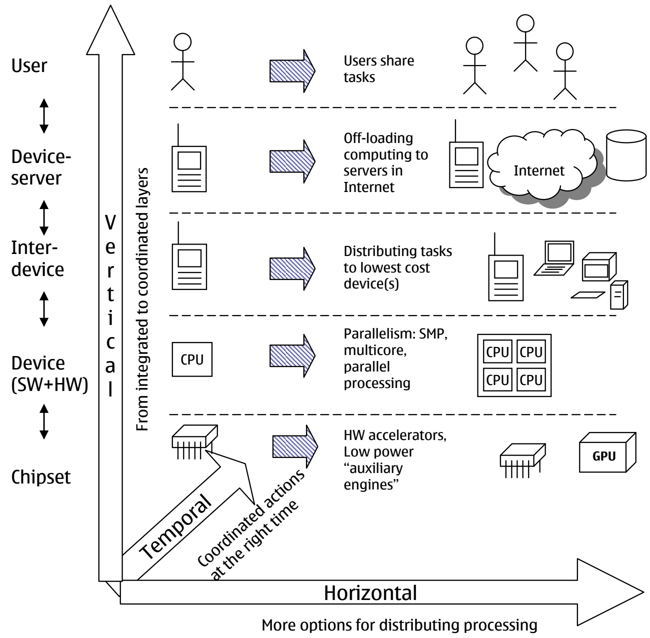

Figure 1: Holistic EPM

One observation is that a mobile device has timers with a wide range of time scales, typically at different vertical levels. For example, application level timers can be order of seconds or minutes while an operating system scheduler works at microsecond scale. These different time scales do not necessarily collaborate much but they may still have negative impact on each other. For example, an application may operate in a bursty mode, but some middleware component may still produce smooth tasks for underlying hardware components, which cannot enter the sleep mode. There is also a connection to horizontal EPM axis. Parallelism means that there are several timers and timing opportunities to combine.

2.4 Holistic EPM Cube

When the three EPM dimensions and their respective EPM mechanisms are connected, we enter the domain of holistic energy and power management. The three dimensions form the holistic EPM cube of Figure 1.

If we first look at the vertical EPM dimension, we can see that various layers have possibilities to be extended in horizontal dimension. For example, at the chipset level a multitude of hardware components can offer concurrent multiprocessing, while the same can be implemented at the software level using a number of parallel threads. At the device level, a number of devices can create a horizontal dimension of collaborating devices for distributed computing. The timing or the temporal dimension is strongly present both with the vertical and horizontal axis. For example, in the horizontal dimension parallel components can work with synchronization, which also has an impact on other levels of the vertical dimensions.

We can make three main observations:

1. EPM dimensions and their energy saving mechanisms can be extended into the other dimensions. For example, with the introduction of multicore processors we can multiply the possibilities for OS level scheduling. Instead of one scheduler in a single, powerful CPU core we have the possibilities for finer granularity energy saving.

2. Various energy saving mechanisms in different places in the three axes can collaborate to enhance their own operation but also the total operation of the device. For example, OS level scheduler may have limited possibilities for trying to allow CPU to enter sleep mode. However, if this scheduling is connected with information about scheduled operations at higher vertical layers (such as application layer deadlines, allowed delays, etc.), the whole situation may be improved.

3. There are possibilities for major trade-offs in order to reduce energy consumption when we have the holistic view of the system. We can, for example, trade a computationally heavy task with lighter communication task. As an example, a heavy local computation in the form of reformatting an image can be avoided by sending the image to an external server that processes the image. Alternatively, a cloud of collaborative devices (smart environment) can allocate a task to the device whose configuration is the most suitable.

3. HOLISTIC EPM OPPORTUNITIES

Typically EPM algorithms handle only a single subsystem at one layer and are often vendor specific. Fast increasing complexity of use cases and devices requires more advanced holistic EPM algorithms that may combine the operation of several subsystems and their own EPM mechanisms. New requirements such as always-on operation, and new user interaction patterns need to be taken into account. At the same time recent changes in industry landscape, including device interoperability, introduction of multi-core processors, and cloud computing introduce new opportunities for EPM. The opportunity with the holistic EPM approach is to effectively combine all three EPM dimensions and sharpening them to collaborate together for the best overall result.

In the vertical dimension the goal is to transition from integrated to coordinated layers, i.e., to make the vertical layers to co-operate in order to optimize the system level energy consumption. A mandatory starting point for this is to have well-defined interfaces towards EPM of upper and lower layers. This way the layers can not only optimize themselves but also offer hooks for other layers to interact, while insulating problems and complexities of one layer.

In the horizontal dimension there are possibilities for concurrent and distributed processing, so this dimension offers more options for mobile systems. The goal is to use parallelism and distribution to transfer functionality to energy-wise lowest cost location. In practice this requires offloading capabilities and trading communication with computing, which means careful analysis and optimization between different possibilities.

The temporal dimension calls for coordinated actions at the right time. With intelligent timing decisions — scheduling, delaying, coupling — the system can fulfill users' needs with minimum energy consumption. The right timing can boost energy saving effects in horizontal and vertical dimensions.

Finally the holistic EPM approach will combine EPM mechanisms at different dimensions. The holistic view will take into account all affecting factors and make the most optimal decisions for the different situations. The decisions may be made at the design or run time depending on the dynamics and required information. In this sense the holistic EPM approach can also be context sensitive and a proper implementation of it can make a notable difference to existing local EPM optimizations.

4. HOLISTIC EPM IN ACTION: ENERGY-EFFICIENT CONTENT DOWNLOAD

In this section we present an example of holistic EPM to make its opportunities more concrete. We will focus on the content download case and illustrate how the different EPM axes can co-operate for energy-efficient downloading of BitTorrent content. We first illustrate the relevant technologies in each EPM dimension and then discuss how these can be combined into a holistic EPM mechanism. Notice that we discuss the different axis in an order that is logical for the example case (and that is different from the previous sections).

4.1 Horizontal EPM - Cloud Computing

Cloud computing aims to provide various functions and services over the Internet by exposing remotely invocable APIs or virtualized resources. The opportunity for energy savings lies in the fact that some functionalities can be transferred from battery-operated device to the cloud. Off-loading work to other devices, which are not energy-limited, naturally saves those resources of the handheld device.

As an example, a recent study has analyzed the energy savings by introducing a server in the cloud for peer-to-peer (P2P) file transfer [3]. This method, CloudTorrent, makes BitTorrent downloads by using a cloud server, which handles the multiple connections which are used to download the content, uploads pieces of content to other peers, and, in general, works as a normal BitTorrent peer. The role of the handheld device is to control the cloud server, keep the user informed about the progress of the download, and, once the BitTorrent download has been completed, move the content to the handheld device for consumption. The principle of CloudTorrent is illustrated in Figure 2.

The key reasons for the energy-efficiency of CloudTorrent are the different service and end-to-end connection speeds of different peers. The slow download bitrate that a mobile BitTorrent client typically experiences, wastes energy.

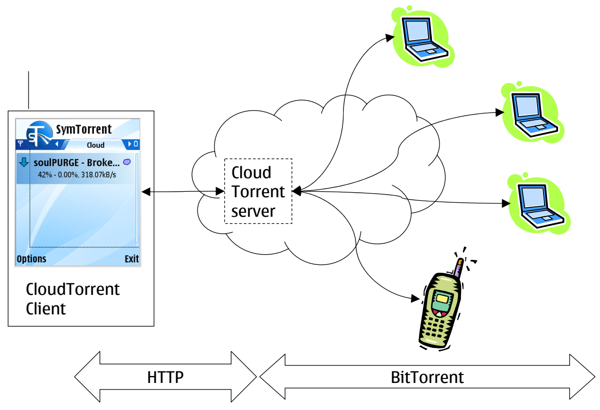

Figure 2: Principle of CloudTorrent.

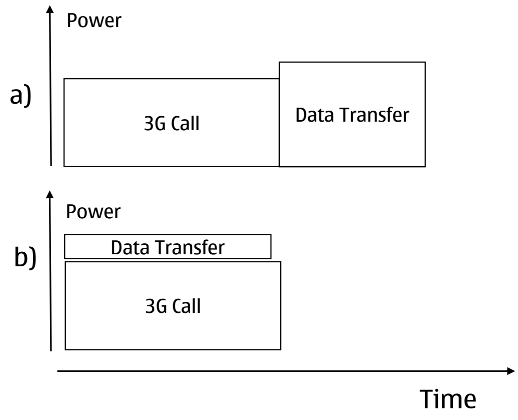

Figure 3: The effect of delayed data transfer. a) Data transfer alone after 3G call takes big amount of energy. b) Making simultaneously data transfer causes just a small extra power consumption.

As data transfer with high bitrates minimizes energy consumption [4], we can save a significant amount of energy by letting the server handle the slow download connections.

In comparison to mobile, the server is able to improve the download speed via these connections because the tit-for-tat mechanism of BitTorrent rewards the higher upload speed of the server with a higher downstream speed. Most importantly when the downloaded content is available at the server, it can dedicate its resources and use its high speed connection to serve a single mobile client with a very high, and thus energy-efficient, bitrate. Minor additional savings arise when computation of hash checks and file formation is performed by the server.

4.2 Temporal EPM - Delayed Data Transfer

The above example of off-loading most of the work to cloud server is already useful for the energy savings. However, we can improve the case further by scheduling the data transfer in a smart way. A recent study [4] shows that major energy savings are possible if the mobile phone performs TCP data transfer when another communication activity, such as cellular voice or VoIP call, is ongoing.

As Figure 3 illustrates, the data transfer is slower but more energy efficient. This observation together with the fact that many data items do not need immediate transfer allows energy savings in the temporal axis. For instance, synchronization of emails, downloading of RSS feeds, or uploading of backups can be delayed by several minutes without sacrificing the user experience. However, this delay may allow them to be scheduled to happen in parallel with voice calls which reduces their energy consumption by 50%-80%.

In the context of the BitTorrent download example this means that instead of transferring the downloaded file from the cloud server to the handheld device immediately when it is ready (or when the relevant parts of it are ready) we would delay the transfer until voice connections are active.

4.3 Vertical EPM- Policy Framework

In [2] we have presented a policy-based approach for system-level power management targeting mobile devices. The proposed framework used policies and context information to allocate resources (such as display backlight or WLAN) to applications as shown in Figure 4. The framework allowed for efficient management of the inherent system-level complexity, but it concentrated mainly on vertical aspects, and did not utilize horizontal or temporal dimension properly.

However, this kind of mechanism allows for moving information between various layers inside the device and to control resources using energy. For example, the Context Monitor in [2] allows to gather information related to the user or the overall status of the device and then Power Policy Manager is used to decide actions that are executed at the low-level system software and hardware.

In case of a BitTorrent download the vertical axis can have a role in deciding when the file should be downloaded. The measurements in [4] show that the energy saving arising from parallel connections depends strongly on the type of the other connections: TCP downloads during 3G voice calls result into 75%-90% energy savings, TCP downloads during VoIP calls result into 30%-40% savings, and TCP downloads when other TCP streams are active in at the same interface result into 0%-20% savings. Therefore the middleware that controls delayed data transfers [5] should not only recognize that a connection is formed but use the attributes of that connection to decide if it reasonable to activate the delayed TCP transfer.

4.4 Holistic EPM Mechanisms

Our example of holistic EPM mechanisms is based on using all the three mechanisms at different EPM axes as described above. This is illustrated in Figure 5. The example starts from the need to move a large file from Internet P2P file service to a mobile device within some time window. For the first phase of the transfer we use CloudTorrent mechanism as described above. In other words, the CloudTorrent proxy in the cloud combines the file streams from various P2P sources until the aggregate file exists. This is the horizon-

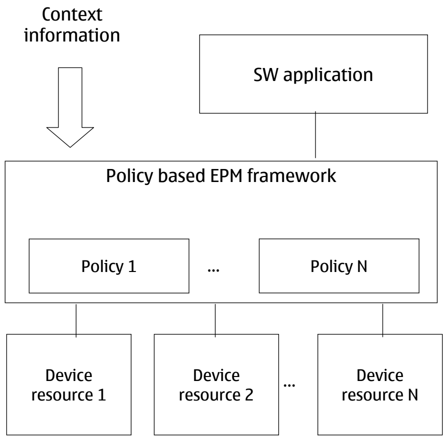

Figure 4: EPM policy framework

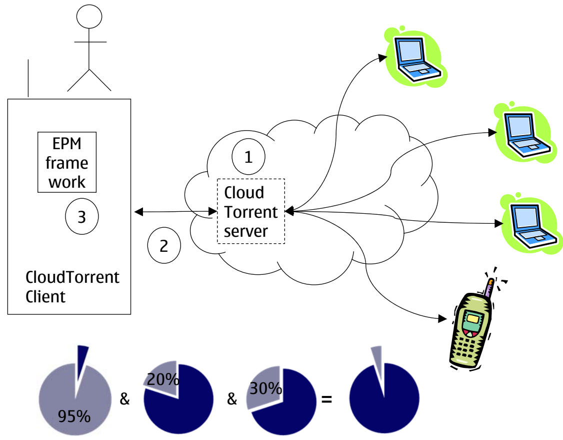

Figure 5: Example of a holistic EPM mechanism.

tal EPM dimension. In the second step we gain additional energy savings by combining this action with a mechanism that delays the transfer of the file from the proxy until a telephone call is made (temporal dimension). Finally, in order to make the decision, the policy-based EPM framework is added with policies implementing the holistic EPM algorithm. The holistic EPM algorithm uses context information (e.g. time, user related issues, networks available) and historical data about telephony usage to evaluate the probability of 3G cellular telephone call with the time window for the file transfer. Thus, we will obtain more energy savings by combining mechanisms at different EPM axes and moving information between them.

The exact savings that the above mechanisms enable depend strongly on the situation. To form a rough estimate, we know that CloudTorrent mechanism can reduce the needed energy to 30% of normal BitTorrent download. Transferring the content via TCP file transfer during a voice call requires around 20% of normal file transfer energy. The combined effect of these two savings would reduce the download energy to around 5% of normal BitTorrent transfer. The mechanisms of the vertical axis, smart selection of download mechanism and time via policy framework, does not necessarily create further savings but, instead, it may improve the user experience by a faster content download. These savings are illustrated by the pie charts in Figure 5.

Note that in the above example the EPM mechanisms at different dimensions influence each other. Most clearly the off-loading created a need for file transfer which was done in the appropriate time with the appropriate mechanisms. However, the interactions are not only sequential but all decisions on all axis influence each other. Offloading the BitTorrent download task to a server may not always be optimal. We need to know what are the costs, in terms of energy, of transferring data in different conditions to make the appropriate decision. The EPM policy framework at the vertical EPM axis would be able to provide such a data for the horizontal axis decision making. Furthermore, the policy framework would also interact with the temporal axis via estimates of different download alternatives which would allow the temporal axis to choose the appropriate download time. A change in context as detected by the policy framework can change the optimal download mechanism and, perhaps, result in a direct download without the server. This case can easily happen if the content is available in near-by devices which can be accessed in a more energy-efficient fashion.

5. DISCUSSION

Finally, there are several other issues involved with holistic EPM approach that need to be considered.

- The user is also a strong part of all EPM dimensions. Obviously, in the vertical dimension the user is the top layer. In horizontal dimension, a group of users may want to collaborate, share data or resources. This is the essence of social networking. Finally, in the temporal dimension, the user ultimately knows how urgent his different needs are and how to balance the contradicting goals. Improvements can be obtained by visualizing holistic EPM to user and by allowing appropriate level of control (if the user wishes).
- User is also the highly unpredictable part of the system. It is often much easier to predict the behavior of technical subsystems. However, there exists a clear demand and challenge for estimating the future behavior of the user at various time scales.
- Too strong focus on EPM could deteriorate other important attributes. While a very holistic view can be good for battery consumption this has to be balanced with developer competence, modularity, maintainability, and other relevant attributes. The goal is to find a proper balance. It would also help the developer if the various trade-offs were more clearly visible.

6. CONCLUSIONS

Traditional EPM mechanisms have challenges managing the fast increasing complexity with requirements for security, speed, connectivity, etc. Simultaneously there are new EPM opportunities with increasing device interoperability, multicore processors, cloud computing, and even improving the energy savings by increasing the developer and consumer awareness of the energy. The opportunity with the holistic EPM approach is to reduce energy consumption and optimize operating times by effectively combining all three EPM dimensions and sharpening them to collaborate together for the best overall result. Holistic EPM approach will combine the mechanisms provided by different dimensions. The holistic view will take into account all affecting factors and make the most optimal decisions for the actual case. Depending on the case, the decisions may be static and made at the design time or dynamic in which case they are made at the run time. In this sense the holistic EPM approach is context sensitive and a proper implementation of it can really improve the energy-efficiency of a mobile device beyond local EPM optimizations. Future work includes application of holistic EPM to key use cases, further quantification of the savings, and mechanisms for holistic EPM optimization.

7. REFERENCES

[1] Advanced Configuration and Power Interface Specification, June 2009. Revision 4.0.

[2] G. B. i Creus and P. Niska. System-Level Power Management for Mobile Devices. In *CIT*, pages 799–804. IEEE Computer Society, 2007.

[3] I. Kelenyi and J. K. Nurminen. CloudTorrent: Energy-Efficient BitTorrent Content Sharing for Mobile Devices via Cloud Services. In *Consumer Communications and Networking Conference*. IEEE Computer Society, 2010.

[4] J. K. Nurminen. Parallel connections and their effect on the battery consumption of a mobile phone. In *Consumer Communications and Networking Conference*. IEEE Computer Society, 2010.

[5] J. K. Nurminen and J. Nöyränen. Parallel data transfer with voice calls for energy-efficient mobile services. In J.-M. Bonnin, C. Giannelli, and T. Magedanz, editors, *MOBILWARE*, volume 7 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 87–100. Springer, 2009.

[6] S. M. Rumble, R. Stutsman, P. Levis, D. Mazières, and N. Zeldovich. Apprehending joule thieves with cinder. In *MobiHeld '09: Proceedings of the 1st ACM workshop on Networking, systems, and applications for mobile handhelds*, pages 49–54, New York, NY, USA, 2009. ACM.