Ultra Low Energy Microcontroller Architectures

Aditya Tandon, *at3g10@sotn.ac.uk, University of Southampton*

***Abstract*—The abstract goes here.**

I. Introduction

THIS report looks into low power and energy techniques for processor architectures

March 17, 2014

II. Asynchronous design

In a field of sensor networks asynchronous processors and microcontrollers have been gaining popularity as they lead to energy efficient designs. The basic principle is that these designs function without a global clock and hence reduce the number of unwanted switching activities in the circuit [1], [2]. To compensate for no clock, the designs usually employ extra hardware for a handshaking protocol [1]. To further conserve energy these designs are event driven [1], [2], [3]. In such a system, the controller is mostly in a state of sleep until it is asked to perform a computation by an event. After performing the task the controller goes back to a sleep state thereby minimizing its active energy [1], [2]. Research has shown that there is no necessary software overhead for these systems due to their event driven nature [1], [2]. This is because these microcontrollers are used for a set of pre-defined tasks and can be simplified. For example, interrupts can be processed as events and there is no overhead to handle concurrent tasks [1], [2]. Designs can be further simplified by employing an in-order design which reduces the amount of hardware and therefore the amount of energy [2], [4].

It can be observed that event driven architectures should have a minimal transition time from a sleep state to an active state. The SNAP/LE architecture addresses this concern by employing an event queue [1]. This resembles a FIFO handler and tasks are executed if there is an event token present in the queue. If there is a token, the appropriate event handler associated with the token is looked up and the task is executed. After executing the task the processor goes into a 'sleep' state if there is no token present [1]. The time taken by the processor to transition between an active and sleep state is the time taken for a token to go through the event queue [1]. The length of the queue can be optimized so that this process is in the order of tens of nanoseconds and therefore this procedure saves energy and is also efficient [1].

Typically these designs can be modularised and stress can be taken off the microcontroller by employing hardware accelerators [1], [2]. Hempstead et al. [2] used the microcontroller only for computational intensive tasks and a separate event processor was employed which was effectively a hard-coded state machine to handle events which required light computation. This reduced the active and leakage power of the main controller. Another type accelerator used was the Message Coprocessor which was responsible for forming and forwarding incoming messages from the radio unit. Timing operations commonly found in wireless applications can be handled by a Timer Coprocessor and therefore can lead to a simplistic, energy efficient implementation for the microcontroller at the expensive of some additional hardware [2]. An architectural implementation of such a system can be seen from Figure 1.

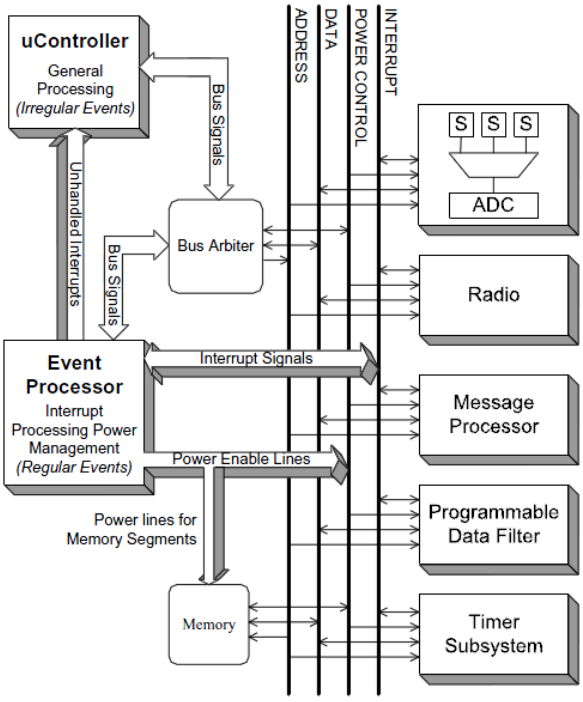

Fig. 1. Event driven design (reproduced from [2]).

The steps taken in desynchronization involve replacing the clock tree with local, asynchronous controllers and converting the flip-flops into latches [3]. This was demonstrated by Neechi et al. [3] where a synchronous AVR microcontroller was used as a template to create an asynchronous version. It was found that the asynchronous controller was about 5 times more energy efficient than the synchronous one [3]. The amount of energy that these prototype processors have taken to execute a particular instruction has been in the range of 10pJ 14pJ assuming an operating voltage of 1.2V and about 2.7pJ/instruction at 0.54V [1], [3], [4].

III. Low power memory

Over the years researchers have come up with different techniques to mitigate energy loss due to memory. A proposal for an adaptive cache for mobile processors could help reduce power [5]. The L2 caches on mobile phones have been found to have access patterns that are not correlated or balanced and therefore there is scope to dynamically adjusting the cache to match the application using it [5]. The compiler does an offline analysis of the application before run-time to determine parameters such as global average miss rates and access rates. Cache access is also monitored during run-time and the run-time information in conjunction with the offline material is used to enlarge or decrease the size of the cache dynamically depending on the need [5]. This proves useful as memory is used much more efficiently. Also, the leakage power is reduced as there are fewer idle cells. This technique was found to give a 13% - 29% reduction in power consumption (using benchmark programs) but there was a small trade-off for speed and area to incorporate this [5].

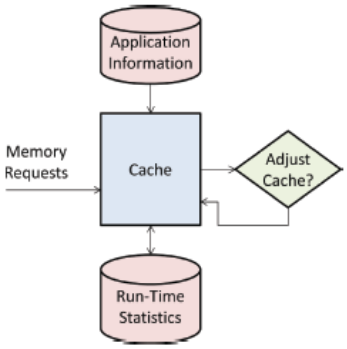

Fig. 2. Adaptive cache design (reproduced from [5]).

Another technique to reduce power is memory compression [6]. Here, a part of the data in volatile memory is compressed in order to reduce the number of logic elements that have to be self-refreshed when a device is turned into a low power state. Rest of the memory can be powered off and therefore the battery life is extended [6]. When there is a request to put a device into a low power state, the memory compression logic takes blocks from a designated memory, compresses it using a compression algorithm and then stores these blocks back to memory. A decompression procedure is followed when the device is in the active state [6]. The compression logic can be implemented in hardware or software and induces some overhead on the battery while performing compression but the power saved by compressing data outweighs this overhead [6].

Non-volatile memories are another area of interest for power reduction and even in performance enhancement [7]. Resistive Random Access Memory (RRAM) is a piece of memory that could be used as a substitute for SRAMs on mobile devices [7]. It was found that an RRAM with a crosspoint structure that uses a diode as a select cell reduces leakage current. This is because the resistance goes up as voltage decreases and the leakage current paths are cut-off [7]. This structure is also area efficient as multi-layered structures can be made [7]. Investigation is still going into this area as peripheral circuit design is harder if RRAMs are used but it could be used as a technique to reduce energy and power [7].

Many other techniques can be used to reduce power. An Intel processor for mobile devices reduces the leakage power in the L2 cache [8]. The data arrays in the cache continue to be in the sleep mode until a Hit signal is generated. Even when there is a hit, only the relevant data array is charged so that it can be activated while the other arrays continue to be in the low power mode [8]. This is a memory partitioning technique and is widely used to mitigate leakage power [2], [8].

IV. Subthreshold design

Devices operating in the subthreshold region show a reduction in energy because the switching activity decreases with supply voltage [9]. However, the propogation delay increases in this region and this gives rise to leakage currents which cause energy dissipation [9]. There exists an optimal operating point in the subthreshold region at which devices can operate. This point can be found through a minimum energy analysis technique in which the energy is plotted against the supply voltage as the voltage is scaled down [9]. Wang and Chandrakasan [9] implemented this idea in a FFT processor to demonstrate how devices can operate in subthreshold regions. Logic elements have to be specifically modified to be catered for subthreshold operation to avoid leakage current. Parallel leakage is a major contributor to leakage current and it occurs when the idle current is comparable to the drive current in circuits [9]. This effect can be mitigated by having reducing or balancing the number of parallel devices in the pull up and pull down path to avoid leakage [9]. Devices should avoid being stacked as this reduces the effective drive current of each transistor in a stack [9].

RAM blocks usually contain six transistor (6T) scheme to enable reading and writing of data. Wang and Chandrakasan have demonstrated that subthreshold conditions make read and write operations harder as they place a sizing constraints on the transistors used [9]. There are also other considerations such as bitline leakage that comes into effect when operating in this region [9]. Therefore, an alternative structure to the RAM is needed to address this problem. In the FFT processor, the RAM uses tristate inverters to create a latch and this is used for write operations. The read operation uses parallel tristate gates and a hierarchical read bitline to mitigate parallel leakage [9].

Therefore, it can be observed that operating in the subthreshold region can help in energy savings but only if the logic is suitably catered for it. Designers might have to construct subthreshold libraries if they want their devices operating in this region. The gains in the long run are satisfactory as it

results in an energy efficient device. The subthreshold FFT chip that was made was found to be "350 times more energy efficient than the low-power microprocessor implementation" which was a microprocessor that did not include subthreshold logic [9].

V. Power management schemes

A. Power management

Chips on mobile phones are now moving towards a multi-core implementation to support the vast functionality that is in demand. The cores are usually heterogenous so that each of them can run at an independent frequency and hence utilized in the best possible way and also reduce dynamic power [10]. A single chip implementation makes it hard for power management schemes to achieve power reduction due to leakage currents. A multi-chip scheme paves way to implementing a partial power off scheme where unused chips can be powered off if they are not in use [10]. This gives rise to the concept of a power domain where different parts of a chip and also different chips can be isolated from each other in terms of power management [10]. Many domains can thus be created and power can be saved. Implementing power domains does pose some problems. For example, the shutdown elements between power domains need to be robust and reliable [10]. μI/Os are used to route signals from domains that might be powered off. These are special circuits used to isolate such signals [10]. To minimize such additional signals, a hierarchical power domain scheme can be adopted which sets a level of precedence for certain domains [10]. So no μI/Os are needed from a higher hierarchy to a lower one as the lower hierarchy cannot be on while the upper one is switched off. Figure 3 illustrates a hierarchical structure that can be implemented on mobile phones.

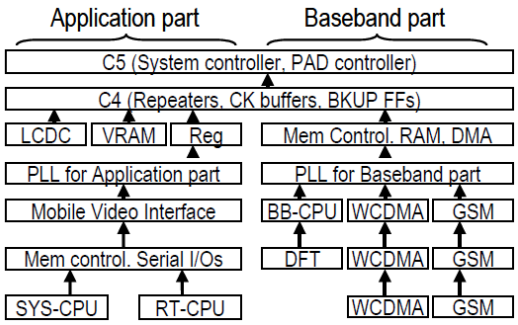

Fig. 3. Hierarchical power domains (reproduced from [10]).

Another problem is the rush current generated while switching on power domains [10]. Hattori et al. [10] described an efficient power switch design to minimize this rush current so that it can be made negligible. They observed very low leakage currents using this design and so the use of hierarchical power domains could be an effective way to save energy [10].

Power gating is another strategy that has been around for quite a while. Here, the processor is switched off due to switches inserted into the power rail [11]. This causes a reduction in power but there are other components, such as state retention registers, power management unit and Low Drop Regulator (LDO) still remain active and contribute to leakage current [11]. To mitigate this, Lueders et al. [11] proposed a scheme based around the LDO. The idea was that a digitally adaptive LDO would drive the micro-controller unit and based on the requirements, it would adapt its drive current and hence reduce power management overhead in low frequency operations [11]. Also, as the LDO is integrated onto the chip, it would be designed with a low output capacitance and therefore take up very little power during sleep and also have a quick transition wake up time [11]. During sleep mode, the LDO can be disabled and it was shown that a power saving of a factor of 4.3 was achieved as compared to power gating [11]. This was easier to implement than power gating as system partitioning was simple and no power switches had to be used [11].

Other power management schemes, include the implementation of different power states in a system [8]. For example, an Intel processor for mobile phones has 6 power states (C0 C6) [8]. The C0 state is the high frequency state and in C6 state the core power is shutdown [8]. The intermediate states involve the power gating of different components such as the core clock, phase locked loops and flushing of L1 caches to reduce dynamic power [8].

B. Dynamic voltage and frequency scaling

Dynamic voltage and frequency scaling (DVFS) is used ubiquitously to improve energy performance in processors [12], [13], [14]. Power is proportional to frequency and the square of supply voltage so scaling down these parameters saves energy [13]. Although delay increases with a reduced voltage so applications that are not time critical can be performed at a lower voltage to avoid performance hits [12]. Multicores on mobile phones can be made much more energy efficient if they ran at an optimum operating point. This operating point is a combination of choosing the correct operating frequency and the number of cores used for an application [12]. Quite often, cores are under-utilized in order to save energy but this actually dissipates more energy as fewer cores are running intensive programs at a higher frequency. Instead of this, more cores can be made available and can operate at a lower frequency [12]. This increases the number of computation resources and also reduces the energy as the operating frequency is low. Also, the work gets executed faster and power dissipation outside the cores can be minimized as these components can be turned off once the computation is finished [12]. Carol and Heiser [12] have come up with a linux governor that actually implements this concept. Frequency is increased if a core is being over-utilized and vice versa. It also disables and enables cores based on their need and this could be a useful tool to incorporate onto devices as an energy saving of upto 25% was observed in this case [12].

Dynamic Thermal Management (DTM) is another important aspect to consider when discussing DVFS [13]. If the temperature of a chip exceeds a certain thermal threshold the frequency has to be scaled down again to prevent the chip from over-heating. Once the temperature goes down the frequency is increased again and this would result in the thermal threshold being exceeded [13]. It can be observed that in some scenarios the operating frequency can oscillate between two frequencies due to DTM and this causes a degradation in power as the chip operates at an unstable frequency [13]. To mitigate this, Kim et al. [13] proposed a DVFS scheme based on an average frequency operating point so that the frequency is stable. Frequencies are sampled on a periodic basis when they exceed the thermal threshold and an average is formed once enough samples are collected [13]. This operating point can be re-sampled for different criteria if, for example, the frequency needs to be lowered or raised. This can lead to a reduction in energy and an energy saving of 12.7% was observed in this scheme [13].

VI. Other techniques

FPGAs have been used to provide quick, cost effective solutions as they have re-programmable capabilities. Yet this re-configurable overhead is also the reason why they consume more power than ASIC designs as power management is more complex [18]. Tuan et al. [18] have investigated low power FPGA applications for battery powered devices and have used a variety of techniques to reduce power. They designed a low power called *Pika* [18]. Voltage scaling was used to scale the core operating voltage to drastically reduce energy. A 1V operating voltage was found to give the best reductions without severely affecting performance [18]. It was found that SRAM cells were a major contributor towards leakage current. Subthreshold leakage was reduced by using a higher voltage threshold (Vt) and gate leakage was reduced by using thicker gate oxides [18]. Though this increases cost and area, the overall energy savings outweigh the cons. Finally, power gating was also used to reduce leakage current [18]. Unused blocks were turned off to save power. NMOS transistors were used as power gates as they are faster. Both NMOS and PMOS were not used in conjunction to save area [18]. Power gating in FPGAs is complex due to the amount of logic that can be gated. Therefore establishing what the smallest block that can be power gated is important [18]. In *Pika*, a tile was the smallest unit that was power gated [18]. A tile here is used to define a configurable logic block (CLB) along with its relevant programmable switch matrix that connects it to other CLBs [18]. The SRAM cells in the switch matrix are not power gated to enable state retention when the rest of the core is powered down [18]. Power gating individual tiles helps in implementing a partial standby mode wherein some logic elements can be powered off and the rest can still remain active. This feature is implemented by having a programmable bit per tile [18]. The overall power savings for all these schemes is illustrated in Table I and it can be seen that a 46% in active power reduction and 99% standby power reduction was observed when compared to a normal FPGA with no power management [18]. There was however a trade-off with performance and area to enable these schemes [18].

Technique	Active Power	Standby Power
Voltage scaling	35%	23%
Low-leakage SRAM config	13%	43%
Power gating	7%	N/A
Standby mode	N/A	51%
Total power reduction	46%	99%

TABLE I
IMPACT OF POWER REDUCTION TECHNIQUES (REPRODUCED FROM [18])

Adaptive body biasing is another technique that was found to be beneficial. It is based on the simple idea that forward body biasing (FBB) decreases the threshold voltage and therefore increases performance and power while reverse body biasing (RBB) increases threshold voltage and reduces leakage current [15]. This adaptive biasing is only applied to certain areas of the chip to suitably alter performance or power and this can give better power savings without affecting performance too much. Gammie et al. [15] have described a tool, *SmartPriMer*, which inserts power management modules into a piece of RTL code. These modules include power domains, adaptive body biasing and other such techniques to reduce power consumption [15]. The tool also generates UPF information for the design that can be used on the later stages for design [15]. These techniques were tested on mobile applications and a 37% reduction in active power was observed [15]. Figure 4 shows the power reduction achieved when the different techniques are used in a 45nm system on chip (SoC) designed for a mobile phone [15]. It can be observed that a mix of Adapative voltage scaling (AVS), RBB and DVFS are used to cut down power when the activity level of a processor is low and least power is consumed when the core is powered down [15].

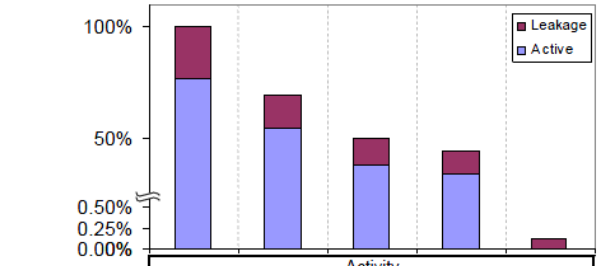

Fig. 4. Power reduction techniques during different processor activities (reproduced from [15]).

Finally, a healthy interaction of hardware and software policies is a good way to reduce power [16]. Most mobile

phones have multi core processors so to make optimum use of these resources parallel computing and task scheduling can be used [16]. Parallel computing involves executing instructions concurrently and an efficient task scheduler can map out a sequence of instructions that can be executed with minimal delay [16]. The use of heterogeneous cores aids schedulers as cores can be catered to different applications and they make good use of the available resources and help in reducing energy as tasks are executed faster [16].

VII. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] V. Ekanayake, C. Kelly, IV, and R. Manohar, "An ultra low-power processor for sensor networks," *SIGARCH Comput. Archit. News*, vol. 32, no. 5, pp. 27–36, Oct. 2004. [Online]. Available: http://doi.acm.org/10.1145/1037947.1024397

[2] M. Hempstead, N. Tripathi, P. Mauro, G.-Y. Wei, and D. Brooks, "An ultra low power system architecture for sensor network applications," *SIGARCH Comput. Archit. News*, vol. 33, no. 2, pp. 208–219, May 2005. [Online]. Available: http://doi.acm.org/10.1145/1080695.1069988

[3] L. Necchi, L. Lavagno, D. Pandini, and L. Vanzago, "An ultra-low energy asynchronous processor for wireless sensor networks," in *Asynchronous Circuits and Systems, 2006. 12th IEEE International Symposium on*, March 2006, pp. 8 pp.–85.

[4] B. Warneke and K. Pister, "An ultra-low energy microcontroller for smart dust wireless sensor networks," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, Feb 2004, pp. 316–317 Vol.1.

[5] G. Bournoutian and A. Orailoglu, "Application-aware adaptive cache architecture for power-sensitive mobile processors," *ACM Trans. Embed. Comput. Syst.*, vol. 13, no. 3, pp. 41:1–41:26, Dec. 2013. [Online]. Available: http://doi.acm.org/10.1145/2539036.2539037

[6] S. Balasundaram, "Increasing the battery life of a mobile computing system in a reduced power state through memory compression," Dec. 20 2007, uS Patent App. 11/450,214. [Online]. Available: http://www.google.com/patents/US20070291571

[7] P. Chiu, P. Lu, and X. Z, "Energy efficiency enhancement in mobile processor memory design using emerging nonvolatile memory," University of Berkley, Tech. Rep., 2013. [Online]. Available: http://www.eecs.berkeley.edu/ pfchiu/EE241_midtermReport.pdf

[8] G. Gerosa, S. Curtis, M. D'Addeo, B. Jiang, B. Kuttanna, F. Merchant, B. Patel, M. Taufique, and H. Samarchi, "A sub-2 w low power ia processor for mobile internet devices in 45 nm high-k metal gate cmos," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 1, pp. 73–82, Jan 2009.

[9] A. Wang and A. Chandrakasan, "A 180-mv subthreshold fft processor using a minimum energy design methodology," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 1, pp. 310–319, Jan 2005.

[10] T. Hattori and et al., "Hierarchical power distribution and power management scheme for a single chip mobile processor," in *Proceedings of the 43rd Annual Design Automation Conference*, ser. DAC '06. New York, NY, USA: ACM, 2006, pp. 292–295. [Online]. Available: http://doi.acm.org/10.1145/1146909.1146986

[11] M. Lueders, B. Eversmann, J. Gerber, K. Huber, R. Kuhn, M. Zwerg, D. Schmitt-Landsiedel, and R. Brederlow, "Architectural and circuit design techniques for power management of ultra-low-power mcu systems," pp. 1–1, 2013.

[12] A. Carroll and G. Heiser, "Mobile multicores: Use them or waste them," in *Proceedings of the Workshop on Power-Aware Computing and Systems*, ser. HotPower '13. New York, NY, USA: ACM, 2013, pp. 12:1–12:5. [Online]. Available: http://doi.acm.org/10.1145/2525526.2525850

[13] J. Kim, Y. Kim, and S. Chung, "Stabilizing cpu frequency and voltage for temperature-aware dvfs in mobile devices," pp. 1–1, 2013.

[14] J. Howard and et al., "A 48-core ia-32 message-passing processor with dvfs in 45nm cmos," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, Feb 2010, pp. 108–109.

[15] G. Gammie and et al., "A 45nm 3.5g baseband-and-multimedia application processor using adaptive body-bias and ultra-low-power techniques," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, Feb 2008, pp. 258–611.

[16] R. Ramirez, E. Rubio, and A. Viveros, "Energy consumption in mobile computing," in *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on*, March 2013, pp. 132–137.

[17] C. H. K. van Berkel, "Multi-core for mobile phones," in *Proceedings of the Conference on Design, Automation and Test in Europe*, ser. DATE '09. 3001 Leuven, Belgium, Belgium: European Design and Automation Association, 2009, pp. 1260–1265. [Online]. Available: http://dl.acm.org/citation.cfm?id=1874620.1874924

[18] T. Tuan, S. Kao, A. Rahman, S. Das, and S. Trimberger, "A 90nm low-power fpga for battery-powered applications," in *Proceedings of the 2006 ACM/SIGDA 14th International Symposium on Field Programmable Gate Arrays*, ser. FPGA '06. New York, NY, USA: ACM, 2006, pp. 3–11. [Online]. Available: http://doi.acm.org/10.1145/1117201.1117203