Estratégias para a Redução de Consumo de Energia e Aumento de Confiabilidade em IoT

Ricardo Reis

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

reis@inf.ufrgs.br

Abstract. *The Internet of Things (IoT) demands new challenges for the design of computing and electronics components. One of the challenges is the power reduction of this large network of connected devices, where the majority is permanently connected. Another important issue, in a large set of applications, especially on critical areas as heath and transport, is reliability. This paper shows an overview of design strategies that we have developed to reduce power consumption and to increase reliability in circuits that are components of the IoT, as reduction of the number of transistors in IoT devices, using optimization techniques and physical design tolerant to radiation effects.*

Resumo. *A Internet das Coisas (IoT) demanda novos desafios no projeto dos dispositivos computacionais e eletrônicos. Um destes desafios é a redução de consumo dos componentes desta grande rede de dispositivos conectados, sendo que a maioria permanece em conexão permanente. Outro aspecto importante, em um grande número de aplicações, especialmente em áreas críticas como saúde e transporte é a confiabilidade. Este artigo visa dar um panorama de estratégias de projeto que temos desenvolvido para a redução de consumo e aumento de confiabilidade de circuitos componentes da IoT, tais como redução do número de transistores nos dispositivos, aplicando técnicas de otimização, novas arquiteturas e projeto físico tolerante a efeitos de radiação.*

1. Introdução

O aumento de crescente de dispositivos conectados na internet das coisas é um dos motivos pelo crescente aumento no número de transistores produzidos anualmente no mundo. A Figura 1, baseada em [SIA 2005], mostra o número de transistores fabricados anualmente no mundo ano a ano. Este crescimento impressionante é devido a 3 fatores principais: aumento do número de transistores integráveis em um chip, aumento do número de produtos que incluem chips embarcados e aumento do número de exemplares fabricados de cada produto. O custo de fabricação de um transistor é relativamente barato. Em [The Economist 2010] é apresentado uma comparação entre o custo de um grão de arroz com o custo de um transistor. O custo de um grão de arroz pode ser equivalente ao custo de fabricação de mais de 125 mil transistores. Isto indicaria que não há necessidade de economizar o número de transistores em um projeto, já que o custo deles é relativamente pequeno. Porém o custo da energia necessária para a operação de um transistor é cada vez mais elevado. Também temos de considerar que um alto consumo de potencia pode reduzir a vida útil de um sistema, assim como aumentar os efeitos de variabilidade que podem provocar um mau funcionamento de um sistema integrado e/ou reduzir sua vida útil. Com a conexão crescente de dispositivos eletrônicos e computacionais na internet, ou seja, na era da internet das coisas, os

problemas de consumo tendem a se agravar, e muito. Portanto, a palavra chave na internet das coisas passa a ser **otimização**, especialmente a otimização de consumo, que deve ser tratada em todos os níveis de projeto de um sistema computacional ou eletrônico. Uma computação sustentável demanda uma otimização em todos os níveis de projeto de um sistema computacional ou eletrônico.

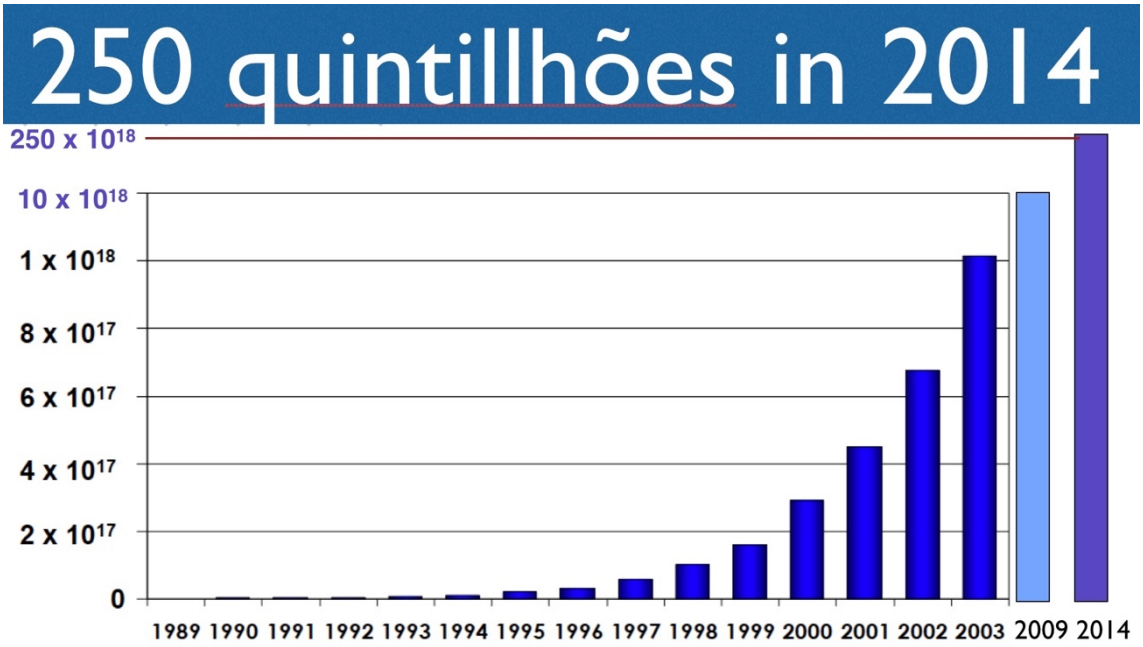

Figura 1. Número de transistores produzidos anualmente no mundo [adaptado de SIA 2005]

2. Internet das Coisas

O termo Internet das Coisas já deu origem a diversos outros termos, como Internet da Saúde (IoH – Internet of Health), Internet das Pessoas (IoP – Internet of People) e Internet de Tudo (IoE, Internet of Everything). Na realidade, este último termo passa a ser o mais abrangente, mas cada um dos demais tem algumas características específicas. Quando se fala em Internet da Saúde, que inclui o monitoramento em tempo real das condições clínicas de uma pessoa, assim como o monitoramento de equipamento implantados em uma pessoa, a questão de confiabilidade é uma questão primordial. E confiabilidade também está relacionada ao consumo, na maioria dos casos. Quando se fala em internet das pessoas, a questão de segurança e privacidade das pessoas tem uma grande relevância. Mas em todos os casos, cresce mais e mais a importância em otimizar o consumo de energia.

Quando tratamos de otimização, significa, que os sistemas integrados devem ser, cada vez mais, dedicados à aplicação prevista, de forma a otimizarem o número de componentes, ou seja, o número de transistores. Outra estratégia importante visando a otimização é o projeto conjunto de hardware e software, onde pode-se gerir o compromisso entre desempenho, consumo e confiabilidade.

Os dispositivos conectados à internet das coisas (ou internet de tudo), podem ter complexidades muito diferentes. Se analisarmos a complexidade quanto ao número de

componentes, podemos encontrar dispositivos pequenos com poucos transistores e dispositivos grandes com bilhões de transistores, como smartphones com SoC avançados (como será exemplificado mais adiante), ou sistemas de controle de veículo de transporte. Evidentemente que os dispositivos grandes vão consumir muito mais energia, mas temos de considerar que a maioria dos dispositivos na internet das coisas são dispositivos com um baixo número de transistores, mas por serem encontrados em grande quantidade, podem representar um consumo total mais importante do que o consumo dos dispositivos ditos grandes. Portanto, a otimização de consumo deve ser efetuada tanto em dispositivos grandes quanto em dispositivos pequenos que estão presentes em grande quantidade. Outro aspecto a considerar é que alguns dispositivos demandam a aplicação de técnicas de aumento de confiabilidade (como os relacionados a sistemas de transporte ou de saúde), que podem aumentar o número de componentes, enquanto outros dispositivos não são críticos, como câmera fotográfica ou de vídeo, em que um erro na visualização de um pixel da imagem não causa maiores problemas.

A Figura 2 [The Connectivist 2014] mostra uma estimativa do número de dispositivos conectados na internet desde 1992, quando eram cerca de 1 milhão de dispositivos, até 2020 quando é estimado que haverá mais de 50 bilhões de dispositivos ligados na rede, sendo que atualmente existem cerca de 35 bilhões de dispositivos conectados. Em [Ihsmarkit 2018] é apresentado o número de dispositivos conectados à rede em 2018, por setores industriais e comerciais, onde quase a metade é na área de comunicação. O crescimento expressivo do número de dispositivos ligados na internet, tem naturalmente provocado um crescimento importante da energia consumida na internet das coisas. Até quando teremos energia para atender esta demanda crescente? Portanto, faz-se necessário o uso de técnicas para diminuir ao máximo o consumo de energia de cada dispositivo ligado na internet das coisas.

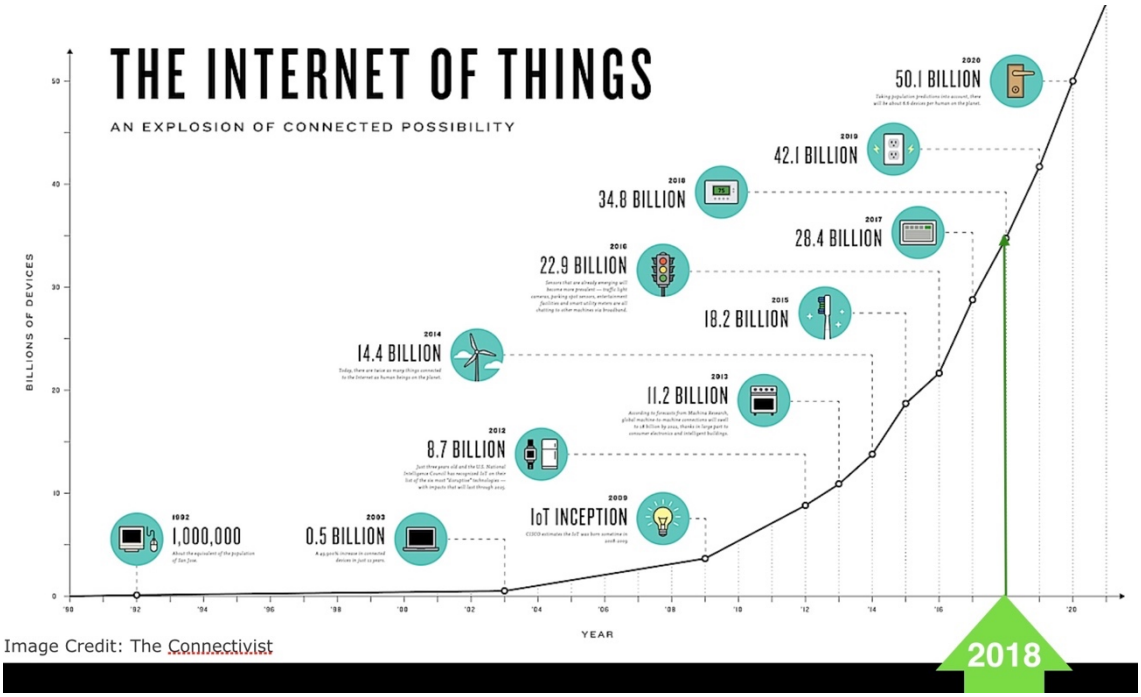

Figura 2. Número de dispositivos conectados na Internet [adaptado de The Connectivist 2014]

Em áreas críticas, como no projeto de dispositivos (chips) implantados em seres humanos (Figura 3), a confiabilidade dos sistemas implantados é evidentemente fundamental. Algumas das técnicas usadas são baseadas na triplicação de circuitos e na análise temporal da propagação de um sinal. Antigamente o projeto de circuitos tolerantes a falhas provocadas por radiação era essencialmente em circuitos que iam para o espaço. Com a redução do valor da tensão de alimentação de circuitos integrados, atualmente, mesmos os circuitos integrados para uso no nível terrestre são sensíveis a erros provocados pela radiação incidente na terra. Portanto, em áreas críticas como em chips implantados em seres humanos, é necessário implementar técnicas de tolerância a efeitos de radiação [Velazco 2007].

Além disso, há o efeito de "*aging*", ou seja, o envelhecimento do circuito, que é mais eminente nas tecnologias nanométricas [Vasquez et al 2012]. Um dos efeitos mais importantes é conhecido como NBTI ("*Negative Bias Temperature Instability"*) que altera a tensão de limiar (*threshold)* dos transistores PMOS, degradando o funcionamento do transistor. Outro efeito que provoca falhas em circuitos ao longo de sua vida é o efeito de eletromigração, que pode provocar curto circuitos ou rompimento de conexões. Para aumentar o tempo de vida dos chips é necessário o uso de técnicas de projeto físico que diminuem a probabilidade de ocorrer eletromigração [Posser 2017]

Figura 3. A implantação de Sistemas em Chip em seres humanos demanda confiabilidade e ultrabaixo consumo

3. Ferramentas de EDA (Electronic Design Automation)

O uso de ferramentas de EDA é fundamental para a otimização do consumo de energia e aumento de confiabilidade. Na Figura 4 podemos visualizar a planta baixa de um circuito integrado, onde as cores mais quentes mostram regiões (*hot spots*) com um

maior consumo de energia, mostrando que em alguns pontos existe uma concentração significativa do consumo. Uma maneira de tratar o problema é modificar o posicionamento das células lógicas no circuito, de forma a distribuir melhor sobre toda a área do circuito as células com maior consumo de energia. Mas isto deve ser efetuado sem comprometer as especificações de área e de frequência de funcionamento (muito depende do roteamento). Outra maneira é diminuir o número de transistores, pois o consumo estático está relacionado com o número de transistores [Reis 2011A].

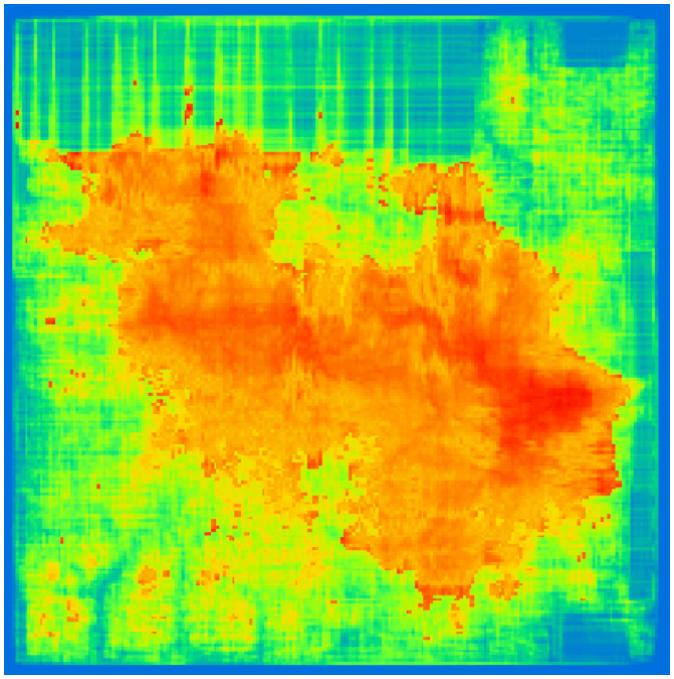

Figura 4. Visualização de densidade de consumo em um chip

4. Redução de consumo através da redução do número de transistores

A redução do consumo de um sistema em um chip é função de uma soma de técnicas e estratégias de projeto aplicadas em diferentes níveis de abstração da concepção de um sistema integrado [Reis 2010]. O somatório dos ganhos é que vai definir o ganho total. Quanto tratamos da síntese física de um sistema em um chip, uma das técnicas é a **otimização** do número de componentes, ou seja, do número de transistores. Na Figura 5 [Reis 2011A] podemos observar duas soluções para a implementação de uma mesma equação. A primeira solução faz uso de 4 portas lógicas básicas (3 portas NOR de 2 entradas e um inversor CMOS), usando um total de 14 transistores. A segunda solução faz uso de apenas uma porta lógica, que executa a mesma função, mas com apenas 8 transistores. Ou seja, a segunda solução, por ter uma redução do número de transistores, também terá um consumo estático proporcionalmente menor. O valor percentual da redução do consumo de potência vai depender do nó tecnológico em que o circuito será implementado. Quanto menor o nó tecnológico, maior tende a ser o consumo estático, devido ao aumento da corrente de fuga. E também varia em função do tipo de tecnologia utilizada, Bulk CMOs, FinFet ou FDSOI, por exemplo. Além disto, no exemplo da Figura 5, podemos ver que a primeira solução possui também 3 conexões entre as portas básicas (e, portanto, também vias e contatos) que são eliminadas na segunda opção com apenas uma porta lógica. Esta eliminação de conexões é cada vez

mais importante, porque diminui o número de conexões a serem implementadas usando as diferentes camadas metálicas. A diminuição do número de conexões diminui a densidade de conexões e, portanto, aumenta a roteabilidade do circuito e contribui também para diminuir o comprimento médio das conexões, o que implica em uma redução do atraso, pois nas tecnologias modernas o atraso em conexões é tão ou mais importante que o atraso no chaveamento das portas lógicas. Um maior espaçamento entre as conexões, também contribui para um aumento da confiabilidade, devido, por exemplo, à redução da possibilidade de eletromigração, conforme já citado anteriormente.

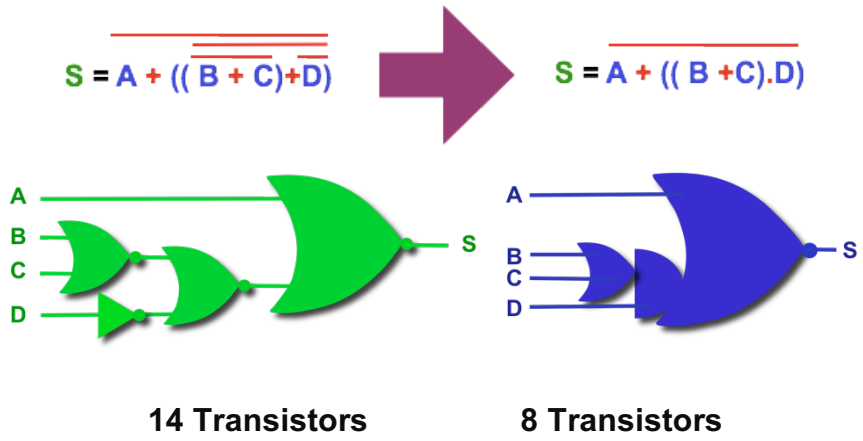

Figura 5. Duas opções para a implementação de uma mesma função [Reis 2011A]

A redução do número de transistores passa pela utilização de ferramentas de EDA (*Electronic Design Automation*) eficientes que efetuem a transformação das equações lógicas de um sistema de forma que além de corresponderem a equações mapeáveis em portas CMOS, façam um uso otimizado de portas lógicas complexas. Em [Conceição 2016] apresentamos uma ferramenta visando reduzir o número de transistores de circuito através da fusão de redes de transistores que apresentam fan out igual a 1. Além disto é fundamental o uso de uma ferramenta de síntese automática de leiaute que consiga efetuar a realização física de qualquer função lógica. Não adianta efetuar uma otimização lógica, se depois for necessário mapear (transformar) as equações em função das portas lógicas disponíveis em uma biblioteca de células tradicional [que não possui poucas funções], como ainda é efetuado nos sistemas comerciais de EDA. Com este objetivo, temos desenvolvido ferramentas de síntese automática de leiaute, como o ASTRAN [Ziesemer 2015] (Figura 6), que permite a geração automática do leiaute de qualquer rede de transistores [Reis 2011].

Outra técnica para a redução de consumo é através do dimensionamento dos transistores. As tecnologias modernas de fabricação de circuitos integrados apresentam um aumento expressivo do consumo estático que chega a ser muitas vezes maior do que o consumo dinâmico. Uma maneira de mitigar o consumo, especialmente o estático é efetuar um dimensionamento de transistores visando otimizar o consumo. Em [Reimann 2016] são obtidas diminuições importantes do consumo através de ferramentas de dimensionamento automático de transistores, também denominadas de seleção de células.

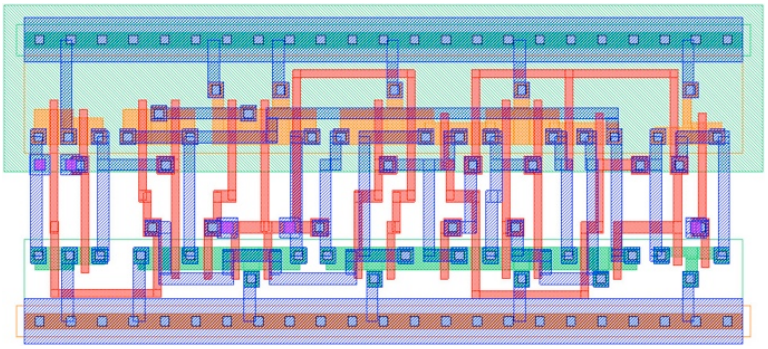

Figura 6. Duas opções para a implementação de uma mesma função [Ziesemer 2015]

5. Confiabilidade

Assim como na redução de consumo, no projeto de sistemas críticos, devemos usar técnicas de aumento da confiabilidade em diversos níveis de abstração. No nível arquitetural, uma técnica muito aplicada é a redundância de módulos, especialmente redundância tripla de módulos (TMR) [Kastensmidt 2006]. Outra é a redundância temporal [Nicolaidis 1999] onde um sinal percorre dois caminhos um com retardo e outro sem retardo, retardo este que deve ser maior do que o tempo de vida de um transiente. A comparação do sinal após percorrer os dois caminhos indica se houve a propagação de um transiente ou não. No nível físico podemos aplicar diferentes técnicas para reduzir ou evitar problemas como eletromigração [Posser 2015] No exemplo da Figura 7, a posição do pino de saída no centro (ponto 4) aumenta o tempo de vida do circuito pois permite reduzir a densidade máxima de corrente nos segmentos da camada metálica.

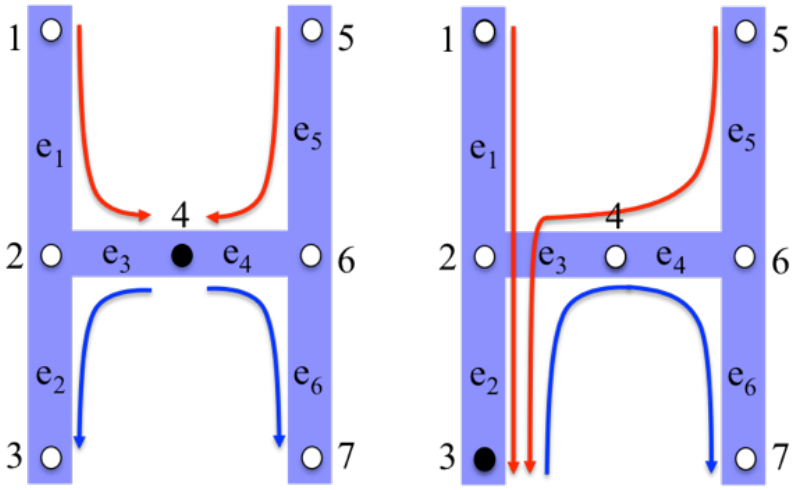

Figura 7. Mudança de densidade de corrente com a mudança de posicionamento do pino de saída [Posser 2015]

Em [Velazco 2007] é apresentado uma série de trabalhos visando mitigar efeitos de radiação em circuitos integrados. Em [Kastensmidt 2006] [Neuberger 2014] [Gennaro 2017] [Aguiar 2016] [Lazzari 2011] [Reis 2011B] são apresentados alguns

dos resultados que obtivemos no desenvolvimento de técnicas visando o projeto tolerante a falhas devido a transientes como os decorrentes de efeitos de radiação.

6. Aceleradores de Hardware

A evolução das arquiteturas de computadores, leia-se hoje, arquiteturas de microprocesssadores tem sido significativa. Na década de 70, um argumento de marketing dos produtores de microprocessadores era o número de instruções que o microprocessador poderia executar assim como a frequência do relógio do microprocessador. Nas ultimas décadas houve uma mudança de paradigma, descontinuando a corrida pelo aumento da frequência do relógio, devido a que o incremento do relógio significa aumento do consumo. Em vez disto houve um incremento do número de núcleos (CPUs) visando o aumento do desempenho. Inicialmente com núcleos homogêneos e posteriormente núcleos de processamento heterogêneos. Muitos dos dispositivos conectados na Internet da Coisas possuem SoCs integrados. Portanto, é importante reduzir o consumo dos mesmos. Um dos dispositivos de grande relevância em IoT são os *smartphones*, que são verdadeiros computadores portáteis, e que possuem SoCs com um conjunto expressivo de funcionalidades e que demandam o uso de técnicas de baixo consumo.

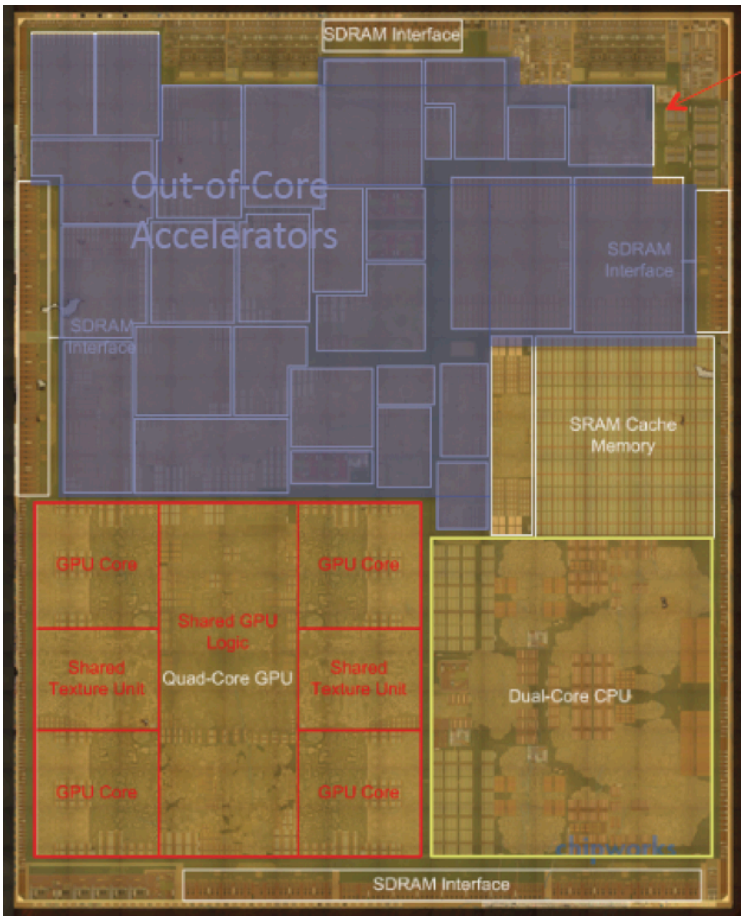

Figura 8. Planta Baixa do Apple 8 com 29 aceleradores de hardware [Anatech, 2014], [SHAO 2016]

Atualmente podemos encontrar chips (sistemas em chip) com várias CPUs e várias GPUs (Processadores Gráficos) na mesma pastilha (como pode ser visto na Figura 8 [Shao 2016] que mostra a planta baixa do microprocessador A8 da Apple). Nesta mesma Figura pode ser observado que cerca da metade da área é ocupada com aceleradores de hardware, que são módulos dedicados à execução de uma função específica. Por exemplo, um módulo de criptografia posicionado junto aos pinos de saída e que vai codificar os dados de saída e decodificar os dados recebidos. Com isto a execução desta função será mais rápida, por ser um módulo dedicado e com apenas o número de componentes para executar aquela função.

Um fato mais importante ainda é que o uso de aceleradores de hardware conduz a uma maior eficiência energética (permitindo uma computação mais sustentável), devido especialmente à redução do número de componentes utilizados para executar uma função. Em um determinado momento, apenas os aceleradores de hardware em uso naquele momento é que são alimentados, ou seja, os aceleradores que não estão em uso, são desconectados da alimentação de energia. Esta estratégia é também conhecida como "Dark Silicon". Podemos até prever arquiteturas compostas essencialmente por aceleradores de hardware, tendo apenas uma ou duas pequenas CPUs para gerenciar os aceleradores de hardware.

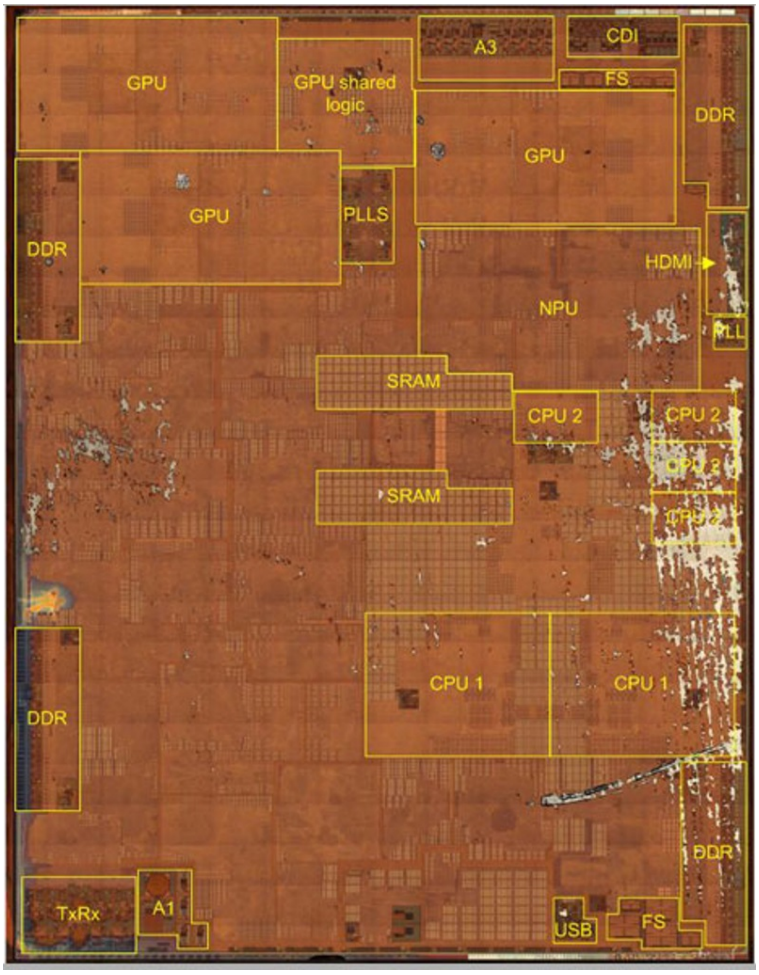

Figure 9. Planta Baixa do Apple 11 com uma NPU [Techinsights, 2017]

Na Figura 9 [Techinsights, 2017] é apresentada a planta baixa do microprocessador A11 da Apple, onde um dos módulos é uma NPU (Neural Processing Unit). A NPU está dedicada essencialmente para o reconhecimento facial [Techinsights, 2017], processando tarefas de aprendizado de máquina de maneira mais eficiente, consumo menos energia do que se fossem realizadas pelas CPUs. As CPUs ocupam cerca de 15% da área do chip e as 6 GPU ocupam cerca de 20%, sendo a maior parte ocupada com aceleradores de hardware. Ou seja, é crescente na arquitetura da linha de microprocessadores da Apple o uso de aceleradores de hardware.

A introdução de uma NPU no A11 é mais um elemento caracterizando a heterogeneidade do SoC (sistema em chip). E podemos esperar arquiteturas cada vez mais heterogêneas, com módulos dedicados para diferentes operações a serem executadas pelo SoC. Com isto podemos esperar dispositivos complexos conectados na Internet das Coisas, mas com um consumo de potência reduzido.

7. Conclusões

Para termos uma computação sustentável, onde é crescente o número de dispositivos conectados na internet das coisas, é fundamental o projeto de dispositivos que sejam **otimizados** em termos de consumo de energia. Atualmente, a maioria dos chips produzidos usam muito mais transistores do que o necessário para executar uma função havendo um espaço significativo para a otimização do número de componentes. Em muitos dispositivos relacionados a aplicações críticas, é também fundamental a aplicação de técnicas visando a tolerância a falhas. Quanto ao consumo, a redução do mesmo deve ser tratada em todos os níveis de abstração em um fluxo de síntese de sistemas integrados, desde a especificação dos mesmos em linguagens de alto nível, até a síntese física. Foram apresentados diversos trabalhos que temos desenvolvido visando a redução do consume e aumento da confiabilidade de sistemas integrados em chip, sendo que maiores detalhes são apresentados nas referências citadas. A palavra chave na era da internet das coisas é **otimização**.

8. Agradecimentos

Agradecemos o apoio do CNPq, FINEP, Fapergs e CAPES pelo apoio financeiro ao desenvolvimento dos trabalhos de nossa equipe, assim como aos alunos de mestrado e doutorado do PGMICRO e PPGC e alunos de Iniciação Científica que tem contribuído com os trabalhos de pesquisa que serviram de base para este artigo.

References

AGUIAR, Y., ZIMPECK, A., MEINHARDT, C., REIS, R. (2016), “Permanent and Single Event Transient Faults Reliability Evaluation EDA Tool”, Microelectronics Reliability, Volume 64, September 2016, Pages 63-67, published by Elsevier B.V., 2016. ISSN: 0026-2714.

ANANTECH (2014), https://www.anandtech.com/show/8562/chipworks-a8

CONCEIÇÃO, C., MOURA, G., PISONI, F., REIS, R. (2017), “A Cell Clustering Technique to Reduce Transistor Count”, 24th IEEE International Conference on Electronics, Circuits and Systems – ICECS2017, Batumi, Georgia, December 5 - 8, 2017, p. 186-189, DOI 10.1109/ICECS.2017.8291996

GENNARO, R., ROSA, F., OLIVEIRA, A., KASTENSMIDT, F., OST, L., REIS, R. (2017), "Analyzing the Impact of Fault Tolerance Methods in ARM Processors under Soft Errors Running Linux and Parallelization APIs", IEEE Transactions on Nuclear Science, Volume: 64, Issue: 8, August 2017, ISSN: 1558-1578, DOI: 10.1109/TNS.2017.2706519

LAZZARI, C., WIRTH, G., KASTENSMIDT, F., ANGHEL, L., REIS, R. (2011), "Asymmetric Transistor Sizing Targeting Radiation-Hardened Circuits", Journal on Electrical Engineering, Springer, DOI10.1007/s00202-011-0212-8, June 2011.

KASTENSMIDT, F., CARRO, L.; REIS, R. (2006), "Fault-Tolerance Techniques for SRAM-Based FPGA", Springer. April 2006, 183 p., ISBN 0-387-31068-1

NEUBERGER, G., WIRTH, G., REIS, R., (2014) "Protecting Chips Against Hold Time Violations Due to Variability", Springer, 107 p., 2014. ISBN 978-94-007-2426-6. DOI 10.1007/978-94-007-2427-3

NICOLAIDIS, M. (1999), "Time redundancy based soft-error tolerance to rescue nanometer technologies". In: IEEE VLSI TEST SYMPOSIUM, 17., 1999. Proceedings... IEEE Computer Society, 1999. p. 86-94.

POSSER, G., FLACH, G., WILKE, G., REIS, R. (2011), "Gate Sizing Minimizing Delay and Area", ISVLSI2011. IEEE Computer Society Annual Symposium on VLSI, Chennai, India, July 4-6, 2011. p. 315-316, ISBN 978-0-7695-4447-2. DOI 10.1109/ISVLSI.2011.92

REIMANN, T., SZE, C., REIS, R. (2016), "Challenges of Cell Selection Algorithms in Industrial High Performance Microprocessor Designs", Integration, Elsevier B. V., Volume 52, January 2016, Pages 347-354, ISSN: **0167-9260**, doi:10.1016/j.vlsi.2015.09.001

REIS, R., (2010) "Redução de Consumo pela Otimização de Componentes", SEMISH 2010, Anais do 37º Seminário Integrado de Software e Hardware, Belo Horizonte, 21 a 22 de julho de 2010, p. 371-379, ISSN: 2175-2761.

REIS, R. (2011A), "Design Automation of Transistor Networks, a New Challenge". IEEE International Symposium on Circuits and Systems, ISCAS2011, Rio de Janeiro, Brasil, May 15-19, 2011. IEEE Press. p. 2485-2488, ISBN: 978-1-4244-9472-9. DOI 10.1109/ISCAS.2011.5938108

REIS, R. (2011B), "Power Consumption & Reliability in NanoCMOS", IEEE NANO, 11th International Conference on Nanotechnology, Portland, USA, August 15-19, 2011 (**invited talk**), p.711-714. ISBN 978-1-4577-1515-0, DOI:10.1109/NANO.2011.6144656

The Connectivist (2014), http://ow.ly/i/5vph6/original

The Economist (2010), 6 de setembro de 2010.

SIA (2015), Semiconductor Industry Association, Rebooting the IT Revolution, disponível em http://www.semiconductors.org/clientuploads/Resources/RITR%20WEB%20version%20FINAL.pdf

IHSMARKIT (2018), IoT Trend Watch 2018, disponível em:

https://ihsmarkit.com/forms/thankyou.html?efid=t+m2jEyFYkJQYyoP3YvuHA==&&gasc_id=862037098&&gasc_label=scrXCLnM7m0Q6siGmwM

Techinsights (2017), http://techinsights.com/about-techinsights/overview/blog/apple-iphone-8-teardown/

VAZQUEZ, J., CHAMPAC, V., ZIESEMER, A., REIS, R., TEIXEIRA, I., SANTOS, M. e TEIXEIRA, P. (2012), "Delay Sensing for Long-Term Variations and Defects Monitoring in Safety–Critical Applications", IN: Analog Integrated Circuits and Signal Processing, Volume 70, Number 2, 249-263, February 2012, Springer, ISSN 0925-1030, DOI: 10.1007/s10470-011-9789-0.

VELAZCO, R , FOUILLAT, P, REIS, R. (2007), "Radiation Effects on Embedded Systems", Springer, June 2007. ISBN 978-1-4020-5645-1

Yakun Sophia Shao (2016), "Design and Modeling of Specialized Architectures, PhD Thesis, Harvard", May 2016. Available at: https://ysshao.github.io/papers/shao2016-dissertation.pdf

ZIESEMER, A., REIS, R. (2015), "Physical Design Automation of Transistors Network", Microelectronics Engineering, V. 148, p. 122-128, December 2015, Elsevier B.V., ISSN: 0167-9317, doi:10.1016/j.mee.2015.10.018