*Sitzungsberichte der Leibniz-Sozietät* *90(2007), 181–199*

Hugo De Man

# Ambient Intelligence: A Giga-Scale Dream Facing Nano-Scale Realities

## 1. The Ambient Intelligence Vision:

Today, we are entering the „embedded everywhere" world in which all objects in our surroundings become intelligent micro-systems interacting with each other and with people through wireless sensors and actuators. The latter also rely on technologies emerging around CMOS, such as 3D packaging, MEMS and polymer displays. Nano-scale biosensors will connect electronics to bio-technology and create new opportunities for healthcare. With IPv6 eve-

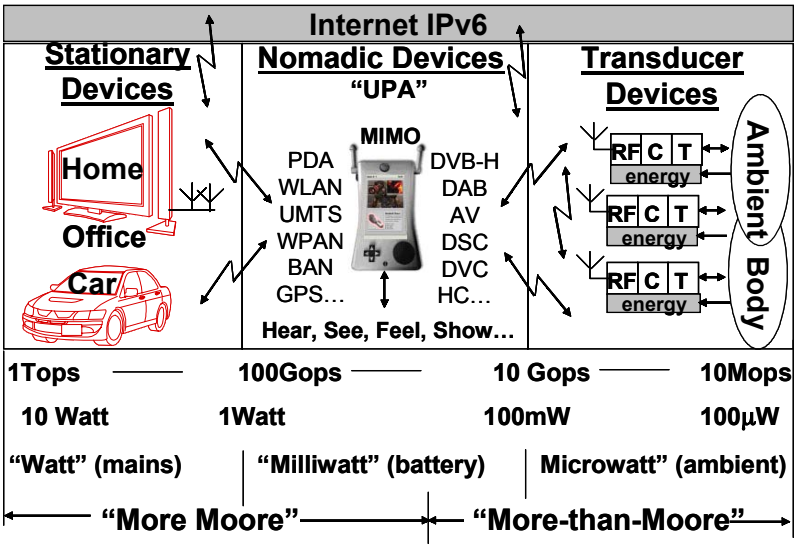


*Fig. 1: Device classes for ambient intelligent systems. UPA: Universal personal Assistant; T:Transducer; C: Signal processing and control. MIMO: multiple input, multiple output multi-antenna system.*

ry object on earth can have its own unique IP address and ultra-low power radio technology will connect all objects by global ad-hoc networking.

This will culminate in the *Ambient Intelligence* (AmI) [1] vision of a world in which people will be surrounded by networked devices that are sensitive to, and adaptive to, their needs.

Fig. 1 is a schematic view of the AmI world. It can be partitioned into three communicating classes of devices depending on their energy sources, cost and computational requirements.

*1. Stationary devices* in the home, office or car require compute platforms that take care of major information processing and execute computational intensive tasks such as interactive audio-visual infotainment in the home or advanced safety, navigation and engine control in cars. Computational power for these applications can reach 1 Tops in the future but packaging and cooling cost limits power for these consumer products to less than 5 Watt (including leakage power). Therefore these devices are called „Watt nodes" [1].

Clearly Watt nodes require a power efficiency of 100…200 Gops/Watt which is 3 orders of magnitude higher than of today's PC microprocessors. However, as AmI components are embedded systems, they need not be general purpose programmable, but must be „just flexible enough" within the intended set of applications. This must be exploited to reduce power.

*2. Nomadic devices.* Every person will carry a Universal Personal Assistant (UPA) powered by battery or fuel cell. It will have natural human interfaces (pen, speech, video, gestures, goggles…). It will provide adaptive multimode wireless and broadband connectivity to the web, to the personal space and to a Body Area Network (BAN) for health monitoring, security and biometric interaction. Natural interfacing requires full multimedia capabilities that automatically adapt the Quality of Service (QoS) to the communication channel, to the computing resources available, and to the attention span of the user. Therefore such a system must be flexible (programmable and reconfigurable).

The wireless part will require a Software Defined Radio (SDR) architecture with multiple antenna (MIMO) techniques for bandwidths in excess of 100 Mbps as well as a configurable front-end to cover all required frequency bands, all of this for a power budget below 500 mW.

The biggest challenge will be in harnessing power in the multimedia and smart audio-visual user interface as well as in the RF power amplifiers. Future multimedia content will be created and consumed by Scalable Video Coding (SVC) [2]. This technique is very error resilient, as needed for wireless chan-

nels, and allows for excellent QoS control. However, computational complexity of SVC decoding and encoding is respectively 9 and 36 times more complex than MPEG-4 for the same resolution. As a result, at least a ten times higher power efficiency will be needed compared to today's known solutions. Peak computational power of an UPA is between 10...100 Gops but peak power for the silicon parts should be less than 1 Watt. Therefore nomadic devices are called the „Milliwatt nodes". They require a power efficiency of the order of 10..100 Gops/Watt.

*3. Autonomous wireless transducers* will be empowered by energy scavenging or lifetime battery. They observe and control our surroundings and form ad-hoc networks communicating with the above two device classes. Today, energy scavenging is limited to 100 microwatt/cm$^3$. So, in spite of the low duty cycle (< 1%) and low bit rates (1 bps…10 kbps), these devices must provide sensing, A/D conversion, computation (< 10 Mops) and RF communication for less than 100 microwatt average, and with microwatt level standby power. Therefore these devices are called „Microwatt nodes".

The digital parts of Watt and Milliwatt nodes demand for „More-Moore" i.e. further relentless CMOS scaling provided we can manage the huge NRE design cost and as long as it leads to further reduction of power and cost per function. Smart autonomous transducers and Milliwatt wireless systems, on the other hand, demand for „More-than-Moore". The challenge is in finding novel combinations of technologies above and around CMOS, for the design of ultra-low power, ultra-simple and ultra-low cost sensor motes for AmI. In the next section we focus first on the challenges for the digital Watt and Milliwatt nodes.

## 2. „More-Moore": Managing Giga-complexity

Watt and Milliwatt devices are consumer products. Unlike general-purpose processors they are not designed for raw performance but for two-orders-of-magnitude lower power for a given task set, at one twentieth of the cost. The computational bottleneck is usually in the memory-intensive digital signal processing parts for multi-mode SDR and processing of multimedia streams. Up to 70% of chip area consists of embedded memory, which is responsible for most of the power dissipation [5].

The battle of standards for 4 G radio and media content processing, and rapidly increasing NRE cost, make ASIC design no longer an option. Instead, energy efficient platforms are needed that can be adapted to new standards

and applications, by preference by loading new embedded system software or by fast incremental modifications to obtain derived products.

Fig. 2 shows that, for the digital Watt and Milliwatt nodes, two major gaps are popping up between AmI dreams and further nano-scaling.

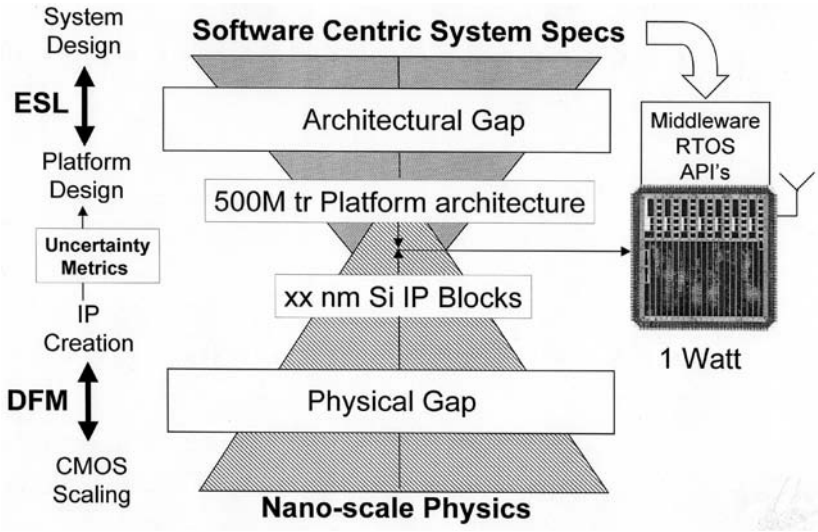


*Fig. 2: Growing gaps between software-centric system specifications and nano-scale physical phenomena affect all levels of the design process and require new skills and industry alliance building. DFM: Design For Manufacturability; ESL: Electronic System Level Design.*

First, there is an architectural gap between AmI dreams and platform architectures. System houses deliver reference specs in C++, MATLAB and the like to semiconductor or fabless companies. These reference specs show functionality without much concern for implementation. The task of platform designers is to map these into energy and cost efficient platforms of hundreds of processors and megabytes of embedded memory. Therefore semiconductor houses have to migrate from pure component manufacturers to domain specialists able to link system knowledge to power-aware nano-scale architectures. Platform design starts far above HDL level and, besides hardware, also a software interface (API) must be delivered as well as a RTOS controlling active and leakage power depending on the platform workload.

To reduce NRE cost, platform design must be based on reuse or EDA supported synthesis of IP-blocks at the processor-memory-bus-periphery level together with the software development environment (compilers, debuggers,

linkers, IS simulators etc.).

Second, there is also a growing physical gap between process technology and platform design caused by nano-scale phenomena such as increased leakage, intra-die variability, signal integrity degradation, interconnect delay and lithographic constraints on layout style. These effects jeopardize the digital abstractions now used for complexity management.

**Managing the architectural gap**

Improving power efficiency of AmI systems by two-orders-of-magnitude requires a rethinking of domain specific compute architectures. Fig. 3 shows the power efficiency PE in Gops/Watt vs. feature size for 32-bit signal processing [6].

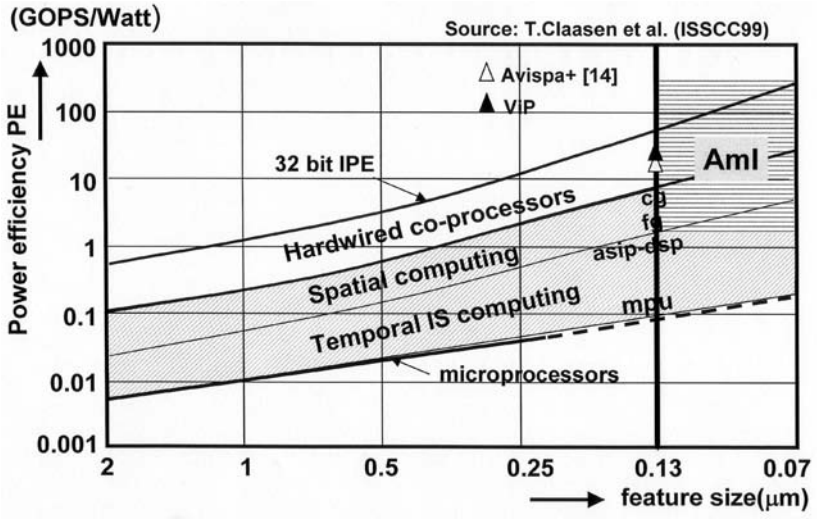


*Fig. 3: Power efficiency vs. feature size for different 32 bit computing architectures. IPE: Intrinsic power efficiency of silicon; ASIP: application specific instruction set computer; fg: fine grain configurable; cg : coarse grain configurable.*

Mono-processors executing temporal computing are the easiest to program, but have the lowest PE. The way to increase computational power is to increase clock speed to the GHz level. This is a poor use of scaling since virtually all additional hardware goes into more levels of cache and in extending instruction level parallelism, which by far does not reach the inherent task and

data level parallelism present in multimedia workloads. As a result we now see a clear trend towards multi-core solutions instead.

Fig. 3 shows that general-purpose microprocessors can dissipate up to 500 times more power for the same task than a pure hardware realization in which all register transfer operations are executed in parallel using distributed local storage. The latter is called the Intrinsic Power Efficiency (IPE) of silicon. The challenge is to approach, as much as possible, the IPE, but for domain specific, flexible, data-dominated platforms.

Spatial computing in Fig. 3 refers to parallel architectures whereby multiple (temporal) processing elements as well as their communication network can be configured at run time under software control. Best PE is obtained if data and instructions are local to the processing elements.

Hardwired multiplexed data paths with local storage and hardware thread control (co-processors) have the highest power efficiency but least flexibility.

Fig. 3 shows that AmI platforms demand for a careful mixture of spatial compute architectures and co-processors with just enough embedded programmability to meet the power budget. It also shows that further CMOS scaling is needed to meet the necessary PE, provided leakage power can be controlled.

We discuss four emerging techniques to reconcile flexibility and power efficiency for data-intensive computing in multimedia and multi-mode digital communication devices. We then discuss the impact of the physical gap.

*1. The need for low power embedded software*

Although power reduction must be performed at all levels of design, the largest gain is at the top. The lack of temporal and spatial locality of data in system level reference code for multimedia and SDR leads to substantial power dissipation. This is especially so for data-dominated systems as most of the power dissipation results from data-transfers to memory. Transfer of data from main memory, cache and local registers costs respectively 10, 5, 2 times as much energy as the actual operation on them by an ALU.

In [7] methods and tools are proposed to transform sequential C or C++ reference code into concurrent SystemC code for low power. These methods help the designer to refine data-types, extract task level parallelism and optimize temporal and spatial locality of data production and consumption This leads to a much better utilization of the memory hierarchy and processor cycles. In addition, tools have been developed to design optimal memory architectures for low power [8] and to minimize cache misses for software controlled caches [9].

These Data Transfer and Storage Exploration (DTSE) techniques have lead to power gains between 2 and 10 on real life designs. Clearly, not paying attention to software transformations for low power leads to a poor utilization of all later architectural optimization efforts.

*2. Exploit parallelism at lowest possible clock speed*

AmI platforms run many concurrent dynamic tasks with widely different sampling rates (audio, video, control, SDR baseband). Hence, a bus-based heterogeneous multi-processor architecture exploiting task level parallelism with localized memory is a natural choice (MIMD). Each processor node must be tuned to the application with just enough flexibility needed by the applications. In each processor, lower levels of parallelism can be exploited (e.g. loop-, data-, instruction-parallelism). This allows for low clock frequencies below 500 MHz, lower supply voltage, use of commercial EDA and Electronic System Level (ESL) design tools and of reuse of large IP blocks. This is mandatory to reduce NRE cost and time-to-market.

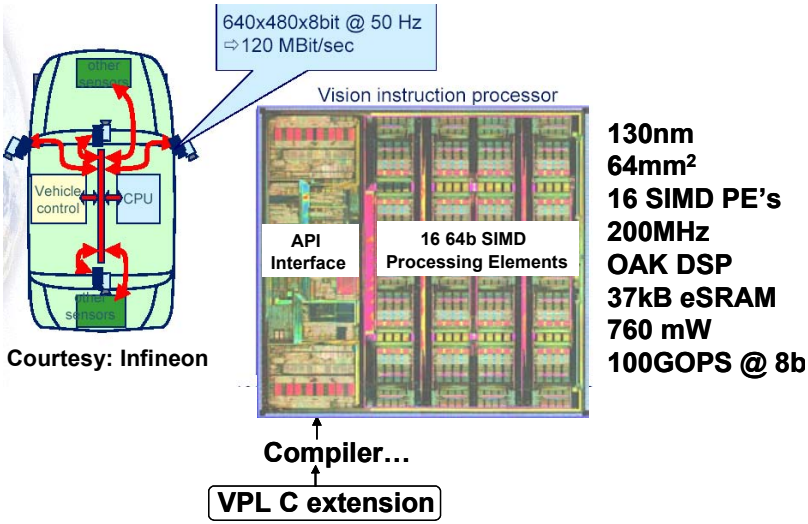


*Fig. 4: Example of a domain specific flexible platform. Infineon Vision Instruction Processor (VIP) for car safety application achieves about 50% of silicon IPE.*

Fig. 4 and 5 show two typical examples of Watt platforms. Fig. 4 is a Vision Instruction Processor (VIP) for car vision from Infineon. Each car mirror has a camera and a VIP that performs real time safety preserving calculations on the image. Notice a network of 16 64 bit SIMD processors each handling e.g.

8*8 bit in parallel. This allows for 100 Gops at 8 bits for a clock frequency of only 200 MHz and 700 mW power dissipation, which corresponds to 38 Gops at 32 bits. Referring to Fig.4 this is close to IPE at 130 nm. An OAK processor allows for programming the VIP for vision applications. It has 204 video oriented instructions.

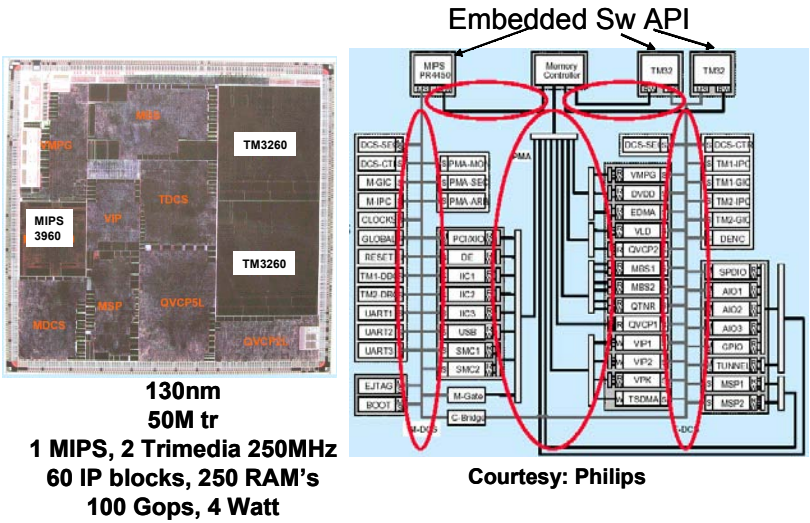


*Fig. 5: Philips Viper2 Bus-Based Nexperia Platform for Advanced Set Top Box applications and digital TV. MIPS controls 60 coprocessors. Two programmable Trimedia VLIW processors allow adaptation to emerging standards.*

Fig. 5 shows the Philips Nexperia platform for high quality digital home video applications handling two video and three audio streams. This 50 M transistor chip in 130 nm technology contains a MIPS processor for global control and 2 VLIW Trimedia processors for adaptation to emerging standards as well as 60 function specific weakly programmable cores for advanced HDTV algorithms. The platform dissipates 4 Watt for a computational power of 104 Gops and local memory bandwidth is 605 MB/s. Exploiting parallelism and locality of data and computation again optimizes the power budget. As a result, the chip contains 250 embedded SRAM's local to the functional units, which again illustrates the need for careful optimization of data-transfer operations.

*3. Exploit task level dynamism*

Advanced multimedia applications are dynamic. Computational power depends strongly on image, music or speech content. Dynamic Voltage/clock

Frequency Scaling (DVFS) [15], driven by the workload, can exploit this. Conventional DVFS is based on a run-time scheduler based on worst-case task profiling in order to meet real time constraints. Recently a Task Concurrency Method [11] (TCM) based on a design time Pareto analysis of tradeoffs between energy and algorithmic tasks has been proposed. Based on the restricted set of Pareto points, a simple run-time scheduler performs an optimal DVFS scheduling on a multi-voltage multiprocessor platform still guaranteeing real-time behaviour. It shows a factor 2 power reduction with respect to traditional DVFS for an MPEG21 graphics application running on 2 StrongARM processors on different supply voltages [8].

*4. Use domain specific processor nodes*

The power efficiency of processor nodes can be substantially improved by defining a specialized instruction set and adding dedicated execution units and memory hierarchy to a processor data path. Application specific instruction set computers (ASIP), VLIW's and run-time reconfigurable 2-D VLIW coarse grain architectures [12] have been successfully used in video applications and SDR baseband processing. Today, design of such processors is greatly facilitated by interactive design systems that automatically generate efficient C compilers, instruction set simulators and HDL descriptions from a description of the processor architecture and its instruction set [13][14]. For example in [14] Silicon-Hive reports the design of a 60 issue VLIW processor (Avispa+) for SDR applications running 5.4 Gops (16 bit) at 150 MHz and dissipating only 150 mW in 130 nm technology. This corresponds to 18 Gops at 32 bits. As shown in Fig. 3, such architectures come close to the PE of ASIC's, but they retain programmability within the application domain and incremental changes are easy to perform from a software specification in C.

## Managing the physical gap: nano-scale realities hit platform architects

Scaling below 90 nm disturbs the digital abstractions and affects both IP design and platform architecting at a time when NRE cost is exceeding 50 M$. The main challenges are:

*1. Gate and sub-threshold leakage power* start to exceed dynamic power. Hi-k gate dielectrics combined with metal- or fully silicided gates (FUSI) will be mandatory to solve the gate leakage problem for high performance applications, but it will take at least until 45 nm before its use in mass production. For AmI applications it is cheaper to keep a thicker gate oxide and, at best, maintain performance by strained silicon.

On the other hand, scaling for performance (low $V_T$, $V_{DD}$) leads to a ten-fold increase of sub-threshold leakage current per technology node. This is unacceptable for AmI applications. A plethora of techniques for leakage control have been proposed [15] but they all affect library design, process technology and even RTOS design and require tight system/technology interaction. Scaling for low leakage (high $V_T$, $V_{DD}$) conflicts with high $I_{on}$ and smaller gate delays. Hence, larger computational power must come from more transistors rather than faster ones. Architecture and technology must be tuned to find an optimum trade-off between clock frequency, degree of parallelism, total power in active mode and leakage power in idle mode.

Leakage power is especially problematic for state retention in SRAM's. Hence $V_T$ scaling in memory cells is hardly possible, and this makes $V_{DD}$ scaling below 1 Volt very challenging although desirable certainly for micro-watt nodes.

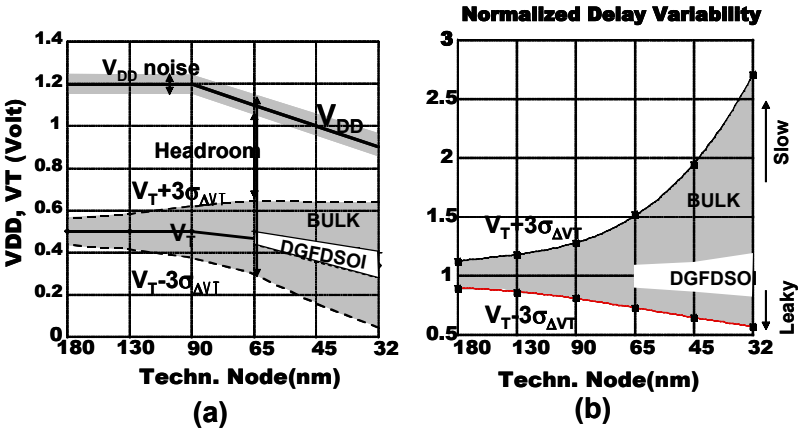


*Fig. 6: (a) Impact of intra-die variability on voltage headroom for Low-Standby-Power bulk CMOS and dual gate fully depleted SOI [17] vs. technology node. (b) Resulting gate delay variability vs. technology node.*

*2. Coping with uncertainty:* nano-level scaling increases the intra-die variability of threshold voltage, drive- and leakage current as they become dependent in the statistical distribution of doping atoms, molecules and photons. These effects can be modeled by the $V_T$ variability $\sigma_{\Delta VT}$. Pelgrom's law states that $\sigma_{\Delta VT}$ = A/sqrt(W.L) which shows its deterioration with scaling. Fig. 6a shows the 3 sigma $V_T$ spread vs. technology node for minimum size, low standby power transistors for both bulk CMOS and metal gate Dual Gate

Fully Depleted SOI (DGFDSOI) [17]. Clearly, voltage headroom $V_{DD}$-$V_T$ (and thus $I_{on}$ and $t_d$) becomes very unpredictable even for neighbouring identical transistors. Fig. 6b shows that gate delay becomes a stochastic variable. This jeopardizes timing closure techniques, reduces parametric yield, requires statistical timing analysis methods and sizing transistors not only for performance but also for yield improvement.

In SRAM's, increasing transistor mismatch prevents $V_{DD}$ scaling below 0.8 Volt during read/ write operation for yield and noise margin reasons especially for bulk CMOS [16]. The use of DGFDSOI can substantially improve this situation from 65 nm on [17]. There is a great need to come up with novel scalable non-volatile RAM with SRAM properties. So, in the coming years, CMOS technology will go through numerous changes that will strongly affect circuit, IP and architecture design.

Platform architects will have to live with this reduced predictability. New methods to design reliable electronics with unreliable components must be developed to avoid worst-case design. One way to do this is by providing a run-time controller that minimizes impact of the variability of the individual system component. Recent work in this direction can be found in [18]. The technique is based on first computing Pareto optimal schedules for a multi-task, multiprocessor architecture. This delivers a discrete set of operating points for a simple run-time task scheduler that guarantees the required cycle budget for minimal power and selects optimal $V_{DD}$, clock frequency and back gate bias for IP blocks [11]. Variability transforms the the nominal Pareto points into point clouds in which the correct point needs to be sampled by on-chip monitors for IP block timing, leakage and temperature. If timing is not satisfied a move is made to the next higher Pareto point satisfying the timing constraints. In frame based processing, monitoring is done per frame period by sacrificing a small part of the million-cycle budget available per frame. This methodology impacts all stages of design and „More-Moore" will critically depend on availability of platform architects skilled in these new design methods.

*3. Lithography challenges:* using 193 nm litho to write features below 90 nm causes increased proximity effects requiring highly regular cell and interconnect architectures to reduce design cost. In addition, Line Edge Roughness (LER) causes about 5 nm variance in line width due to the granularity of resists and phonons. This effect is another contributor to the variability issues mentioned above and needs innovative solutions. One way could be to use nano-technology to assemble atomic layers instead of litho and etching techniques but this is far from exploitation today.

*4. Interconnect challenges:* scaling causes faster logic but slower global interconnect. Chemical polishing techniques cause thickness variability up to 40% depending on the wiring context and strong capacitive interline coupling leads to a poor signal integrity. This will impact the way to design and layout on-chip communication. Global bus structures as in Fig. 5 do not scale well to higher complexity and global synchronism will have to be abandoned in favour of Globally Asynchronous Locally Synchronous (GALS) architectures. Furthermore, in [19] it is shown that dynamic power in standard cell wiring will be 5 to10 times higher than in the cells themselves. In contrast, in [20] it is shown that structured and abutment based data-path layout shows a five-fold higher energy efficiency than standard cell layout at the expense of design time. Hence a renewed EDA activity on automated structured layout generators at nano-scale level is needed if we do not want to loose at the circuit level what we gain at the architectural level.

**Future platforms: towards tile based network-on-a-chip (NoC)**

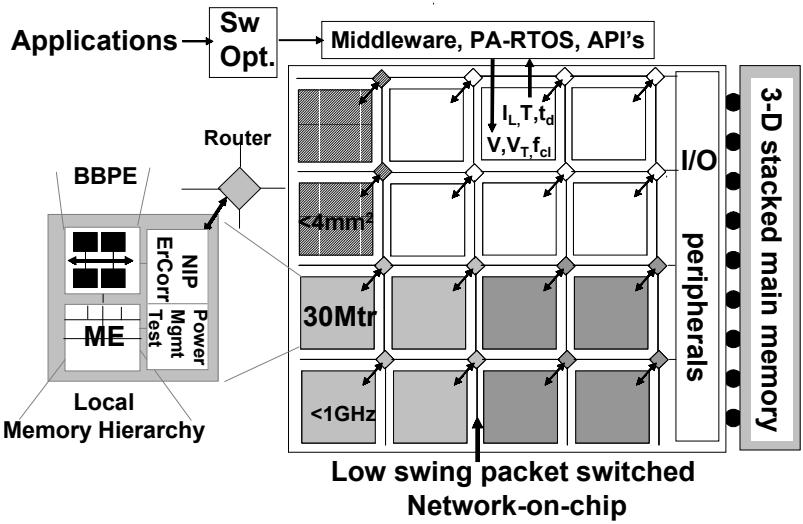


*Fig. 7: Nano-scale platforms will be Globally Asynchronous,Locally Synchronous. Synchronous, Bus-Based-Multi Processors (BBPE) communicate over Networks-on-Chip. Network Interface Processor (NIP) decouples communication and computing. Error-correction (ErCorr) guarantees bit-error rate. Middleware, Power-Aware RTOS and Programmer Interfaces to be co-developed with the platform.*

Based on the above, we can make a conjecture about future platform architectures, illustrated in Fig. 7. At the 45 nm node, even at a modest 500 MHz clock rate, 2 mm local copper wires represent a delay of 30% of the clock cycle. Hence iso-synchronous zones must be restricted to compute tiles of less than 4 $mm^2$ or about 30 Million transistors. Each tile can easily contain a synchronous bus-based multi-processor (BBPE) with local clocking, memory hierarchy, local power control and BIST test processor.

Synchronous tiles communicate globally in an asynchronous way over low swing interconnect to guarantee both low power and ease of design. Since global busses do not scale and interconnect (rather than transistor) becomes the scarce resource, Dally [21] has suggested to „route packets, not wires". This leads to the Network-On-Chip (NoC) concept [22]. Compute tiles are routing data-packets over a structured network fabric of shared wires while a standardized network Interface Processor (NIP) per tile decouples communication and computation. The NIP implements the communication protocol programmed upfront from a software service layer. It hides the network details from the computing tile and thus allows for a plug-and-play design strategy necessary for fast turn-around design of derivative platforms.

In [23] the first silicon implementation of synthesizable NoC IP is presented. A router IP block can be programmed for guaranteed and best-effort services. In 120nm an aggregate bandwidth of 80 Gbit/s in 0.26 $mm^2$ is reported for a 5*5 router.

Low swing global interconnect is advantageous for low power but is more sensitive to supply and crosstalk noise. So, we must learn to live with errors by applying error coding/decoding (ErCorr) techniques that guarantee a bit error rate consistent with the required S/N ratio of the digital signals [24] as is normal practice in communication systems.

Finally power hungry communication to external DRAM memory can be reduced by 3D stacking of DRAM on top of the platform chip [3].

**The devil is in the software**

AmI systems must be adaptive to new software services, standards, communication protocols etc. As a result, not only the hardware must be delivered but also a number of software layers as shown in Fig. 7. First, similar to the JAVA virtual machine concept, a middleware layer is required to insure interoperability. Second, a platform specific, Power-Aware (PA) RTOS is needed to schedule tasks dynamically on the computing tiles and to minimize power consumption at run time. Finally low-level API's are needed for the

processors in the tiles. The development cost of this Hardware dependent Software (HdS) easily exceeds that of hardware development itself and adds considerably to the NRE cost. Therefore, the development of such complex systems will be restricted to a few grand industry alliances that can organize disciplined armies of engineers to design the AmI products of the future. To alleviate this problem, research is needed to come up with efficient methods for co-design of hardware and software platform architectures.

## 3. „More-than-Moore": ultra-creativity for ultra-low power and cost

Platform design of stationary and nomadic devices is all about managing giga-scale system complexity implemented in nano-scale CMOS. In contrast, design of wireless transducer network devices requires creative engineering to get to the ultimate limits of miniaturization, cost reduction and energy consumption.

This leads to the need for „More-than-Moore" i.e. a cost-effective integration of CMOS with MEMS, optical and passive components, new materials, bio-silicon interfaces, lifelong autonomous energy sources and grain-size 3D packaging. The complexity is not in the number of transistors but in combining technologies, circuit- and global networking architectures to obtain utmost simplicity for the sensor nodes themselves.

Microwatt devices are low duty cycle (< 1%), low throughput (1 b/s–10 kb/s) microsystems that unify nearly all design art in one package: sensor, signal conditioning, A/D conversion, signal processing (compression, interpretation, encryption), power-aware MAC layer, picoradio (wake-up Rx, data Tx), antennae, energy management and energy scavenging from the environment.

Fig. 8a shows an IMEC SiP realization of a 1.4 $cm^3$ 2.4 GHz EEC, ECG sensor mote using laminate packaging of bare dies, a solar-cell battery charger and integrated antenna. This system consumes 500 μW at 400 bps and 1% duty cycle. However further integration to e-grain size and lower power will be necessary. Indeed, for true energy scavenging, only solar cells, piezo-electric MEMS (vibrations, shocks) and thermal generators have proven to be succesful but their average power capacity is limited to 100 microwatt/$cm^3$ [25]. This means less than 10 mW peak power during active periods for a 1% duty cycle. So peak power per subfunction should be below 2 mW. For 1% duty cycle and 90nm technology this allows for about 5M 8bit ASIP operations/sec for all data and signal processing and less than 2 nJ/bit transmission energy for 10 kbps. This requires an order of magnitude power reduction with

respect to Zigbee and Bluetooth, especially in the RF part. This can only be obtained by covering very short distances (< 10 m) and/or using multi-hop networking and RF architectures of utmost simplicity. Pioneering work in this direction can be found in [26] [27] (Fig. 8b).

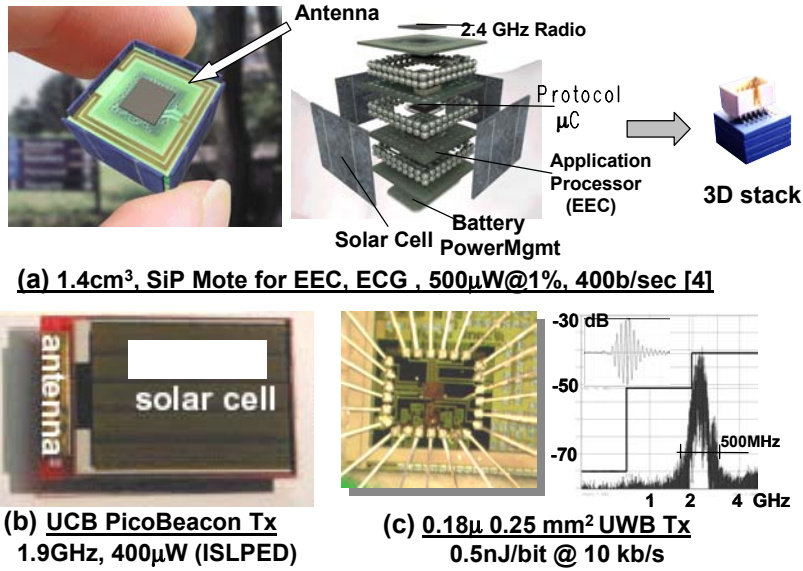


*Fig. 8: Sensor radio's: (a) Example of a sensor mote SiP for wearable medical applications (IMEC); (b) Solar Cell Powered Ultra-Low-Power Transmitter (U.C. Berkeley) (c) Ultra-Low power Ultra-Wide Band transmitter for Body Area Network (IMEC).*

For Body Area Networks and accurate positioning UWB radio could be a solution. Fig. 8c shows an IMEC design of a 180 nm 0.6*0.6 $mm^2$ UWB transmitter in the 3-5 GHz band. All circuits are digital except the power output stage and a triangular wavelet shaper. This extreme simplicity leads to 0.5nJ/bit transmit energy at 10 kb/s and 10 pulses/bit. Active power is 2 mW. In [28] CSEM and Delft University approach UWB by ultra-wide band FM (UWB-FM) which leads to very simple TxRx architectures [28].

Clearly there are great engineering challenges in this domain which is so crucial for AmI. Not the least of the challenges is security of sensor networks which, together with network protocol and data storage and lightweight operating system [29] require most of the computational power. New ideas in efficient, safe and cheap encryption for sensors are urgently needed and will

require ultra low voltage (< 500 mV) low leakage computation such as proposed in [30][31]. In [32] a 32 bit adder in 130 nm reaches 0.3 pJ/add at 300 mV with forward body bias. With such a circuit technique and using massive parallelism, one would be able to reach about 1 Gops/mW which opens great perspectives if we can overcome the architectural and nano-scale challenges mentioned above.

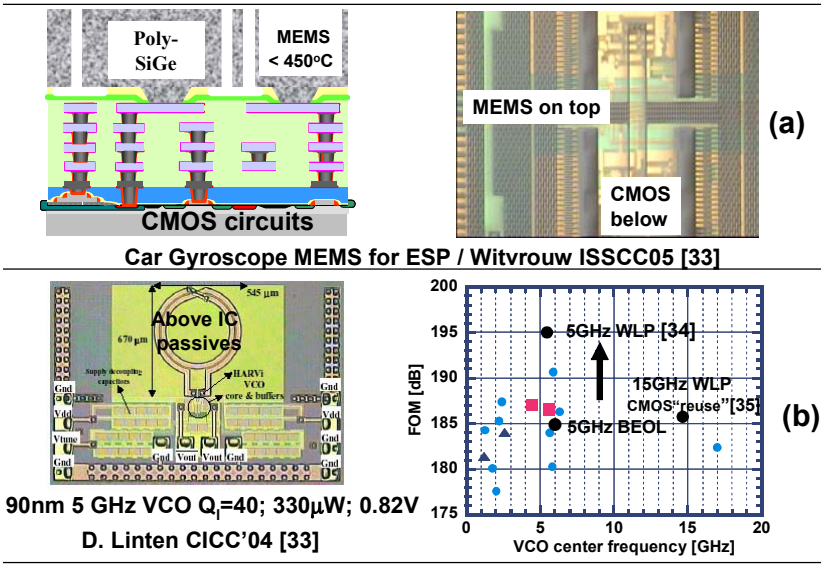


*Fig. 9: Above IC processing (a) Low temperature poly-SiGe MEMS process [33] (b) Above IC Hi-Q inductors and capacitors lead to low power VCO, area savings and reuse of CMOS circuit design.*

Massive deployment of sensors or RFID tags requires cost reduction to the single dollar or cent range but this conflicts with very cheap integration of standard CMOS with non-CMOS devices s.a. passive components and MEMS and even with biotissues. SiP techniques and techniques for low temperature wafer scale integration „above silicon" allow for the „reuse" of silicon wafers and for adding passive components and MEMS on top of them at low cost. In [33] and Fig. 9a an above IC poly-SiGe MEMS technology is reported and illustrated by a MEMS gyroscope on top of the CMOS signal processing.

Fig. 9b shows above IC processing for depositing high quality RF passives [34]. This allows to run a 90 nm CMOS VCO at 0.82 Volts for 330 µW

power and – 155 dBc/Hz @ 1MHz phase noise. In [35] it is shown how „reuse" of the 5 GHz CMOS part with another above IC inductor leads to a 15 GHz oscillator with better figure-of-merit than with state of the art full CMOS integration.

Healthcare and wellness will be an important application domain of AmI and interfacing between electronics and living bio-tissues will be of crucial importance. In [37][38] new breakthroughs in coupling ionics to electronics and in cell manipulation are reported.

## 4. Conclusions

Ambient Intelligence is the next wave of information technology for enhancing human experience. It implies a consumer-oriented industry driven by software from the top, and enabled and constrained alike by nano-scale physics at the atomic level.

„More-Moore" will be needed to deliver Giga-ops computation and GHz communication capabilities for stationary and wearable devices. The grand challenge will be to design flexible multi-processor platforms with two-orders-of-magnitude lower power dissipation than today's microprocessors at one twentieth of the cost, while coping with the realities of nano-scale physics. We have presented a number of emerging techniques to cope with this challenge. Design of such systems will depend critically on our ability to create multi-disciplinary alliances able to cover the huge span between AmI dreams and their implementation in the interaction of billions of nano-scale devices.

On the other hand, „more-than-Moore" technology is needed for the design of autonomous wireless sensor networks. The complexity is not in the number of transistors but in clever combinations of technologies, circuits and system architectures to design ultra-low power, ultra-low cost, ultra-simple sensor motes for AmI.

Both „More-Moore" and „More-than-Moore" rely in the first place on the availability of creative engineers able to approach the problems from a holistic perspective. Engineering schools should reflect on the impact of this evolution on education and research by paying more attention to the application of technology to solve the grand societal challenges of the 21st century.

## References

[1] T. Basten et al., „Ambient Intelligence: Impact on Embedded System Design", *Kluwer Academic Publishers,* 2003

[2] J. Ostermann et al.,"Video Coding with H.264/AVC: Tools, Performance and Complexity", *IEEE Circuits and Systems Magazine,* pp. 8-28, First Quarter 2004
[3] E. Beyne, „3D Interconnection and Packaging: Impending Dream or Reality" *ISSCC Dig. Techn. Papers*: pp. 138-139, Feb. 2004
[4] T. Torfs et al., „Wireless Network of Autonomous Environmental Sensors", *Proc. IEEE Sensors 2004,* Vienna, Oct. 2004
[5] R. Gonzales et al.,"Energy Dissipation in General Purpose Microprocessors", *IEEE J. of Solid State Circuits,* Vol. 31, No.9, pp. 1277-1284, Sept. 1996
[6] Theo Claasen, „High Speed: Not the Only Way to Exploit the Intrinsic Computational Power of Silicon", *ISSCC Digest of Technical Papers,* pp. 22-25, Feb. 1999
[7] F. Catthoor et al., "Code Transformations for Data Transfer and Storage Exploration Preprocessing in Multimedia Processors", *IEEE Design & Test,* Vol. 18, No. 3, pp. 70-81, May 2001
[8] H. De Man et al., „Filling the Gap Between System Conception and Silicon/Software Implementation", *ISSCC Dig. of Tech. Papers* pp. 158-159, Feb. 2002
[9] F. Catthoor et al., „Data Access and Storage Management for Embedded Programmable Processors", *Kluwer Academic Publishers,* 2002
[11]P. Yang et al., „Energy-Aware Runtime Scheduling for Embedded Multiprocessor SoC's", *IEEE Design and Test,* Vol. 18, No. 3, pp.70-82
[12]B. Mei et al., „ADRES: An architecture with tightly coupled VLIW processor and coarse-grained reconfigurable matrix" in *Field-Programmable Logic and Applications*, 2003
[13]see e.g. www.retarget.com; www.coware.com; www.tensilica.com; www.arc.com
[14]www.siliconhive.com
[15]T. Sakurai, „Perspectives on Power-Aware Electronics", *ISSCC Dig. Of Techn. Papers,* pp. 26-27, Feb. 2003
[16]K. Itoh et al., „Review and Future Prospects of Low-voltage Embedded RAMs", *Digest of CICC 2004,* Oct. 2004
[17]M. Yamaoka et al., „Low Power SRAM Menu for SoC Application Using Ying-Yang Feedback memory Cell", *Dig. Symp. VLSI Circuits* pp. 288-291, June 2004
[18]B. Colwell et al, „Better than Worst Case Design". Theme issue in *IEEE Computer,* Vol. 37, No.3, pp. 40-73
[19]C.J. van der Poel et al., „On ambient intelligence, needful things and process technologies", *Proceedings of ESSCIRC 2004*, pp. 3-10
[20]O. Weiss, M. Gansen, and T. G. Noll, „A flexible data path generator for physical oriented design," in *Proc.2001 ESSCIRC)*, pp. 476–479. Sept. 2001
[21]W. Dally et al. , „Route packets, not wires: On-chip interconnection networks", *Proc. Design Automation Conf. 2001,* pp. 684-689
[22]A. Jantsch, H. Tenhunen (EDS.), „Networks on Chip", *Kluwer Academic Publishers,* 2003

[23]E. Rijpkema et al., „Trade-offs in the design of a router with both guaranteed and best-effort services for networks on chip" *Proc. 2003 of DATE Conf.*, pp. 350-355, March 2003

[24]D. Bertozzi et al., „Low Power Error Resilient Encoding For On-Chip Data Busses" *Proc. 2002 DATE conf.* , pp. 102-109, March 2002

[25]S. Roundy et al. „Energy Scavenging for Wireless Sensor Networks with Special Focus on Vibrations", *Kluwer Academic Publishers,* Jan. 2004

[26]J. Rabaey et al., „PicoRadios for Wireless Sensor networks: The Next Challenge in Ultra-Low Power Design", *ISSCC Dig. Of Techn Papers,* pp. 200-210, Feb. 2002.

[27]S. Roundy et al., „A 1.9 GHZ Transmit Beacon using Environmentally Scavenged Energy", *Proc. IEEE ISLPED'03 ,* Seoul/Korea, 2003

[28]J. Gerrits et al., „UWB Considerations for „My personal Global Adaptive Network" Systems", *Proc. Of 2004 ESSCIRC*, pp. 45-56, Sept. 2004

[29]see e.g. http://www.tinyos.net/

[30]A. Wang et al., „A 180mV FFT Processor Using Sub-Threshold Circuit Techniques", *ISSCC Dig. of Tech. Papers,* pp. 292-293, Feb. 2004

[31]B. Calhoun et al., „Ultra-Dynamic Voltage Scaling Using Sub-Threshold Operation and Local Voltage Dithering in 90 nm CMOS", *ISSCC Dig. Of Techn. Papers,* Feb. 2005

[32]K. Ishibashi et al., „A 9 µW 50MHz 32b Adder Using a Self-Adjusted Forward Body Bias in SoCs", *ISSCC Dig. of Tech. Papers,* pp. 116-117, Feb. 2003

[33]A. Witvrouw et al., „Processing of MEMS Gyroscopes on Top of CMOS ICs", *ISSCC Dig. of Tech. Papers,* Feb. 2005

[34]D. Linten et al., „A 328 µW 5GHz voltage-controlled oscillator in 90nm CMOS with high quality thin film post-processed inductor", *Proc. CICC,* Orlando, Sept. 2004

[35]G. Carchon et al., „Thin Film as Enabling Passive Integration Technology for RF-SoC and SiP", *ISSCC Dig. of Tech. Papers,* Feb. 2005

[36]M. Dubois et al., „Integration of High-Q BAW Resonators and Filters above IC", *ISSCC Dig. Of Techn. Papers,* Feb. 2005

[37]P. Fromherz, „Joining Ionics and Electronics: Semiconductor Chips with Ion Channels, Nerve Cells and Brain Tissue", *ISSCC Dig. Of Techn. Papers,* Feb. 2005

[38]H. Lee et al., „An IC/Microfluidic Hybrid Microsystem for 2D Magnetic Manipulation of Individual Biological Cells", *ISSCC Dig. Of Techn. Papers,* Feb. 2005