DATE Panel
Chips of the future: Soft, Crunchy or Hard?

Pierre G. Paulin (organizer and moderator)
SoC Platform Automation,
Central R&D, STMicroelectronics,
Ottawa, Canada
pierre.paulin@st.com

Abstract

Today's electronic products are composed of an increasingly diverse set of IC's, ranging from dedicated ASIC's, domain-specific ASSP's, platform FPGA's, to general-purpose FPGA's. With increasing integration, a mix of different fabrics on a single SoC becomes possible, combining ASIC-style standard cells, embedded FPGA's, mask-programmable sea-of-gates, and programmable processors. The panelists will present their vision of the fabric which will dominate SoC's in 90nm technologies and beyond, based on industrial trends and case studies. They will also outline the key CAD tool challenges for the chosen fabric.

1. Introduction

The continued increase in the non-recurring expenses (NRE) for the manufacturing and design of nanoscale systems-on-chip (SoC), in the face of continued time-to-market pressures, is leading to the need for significant changes to their design and manufacturing. The mask set manufacturing NRE of an SoC has been multiplied by a factor of ten in about three process technology generations, exceeding 1M$ for current 90nm process.

At this cost, many smaller design houses cannot afford the financial risk of a tape-out. For example, for a chip sold at a price of $5, which is close to the center of gravity of the overall semiconductor business, and a profit margin of 20%, this implies selling over one million chips simply to pay for the mask set NRE. This does not even account for the accompanying increase in design NRE (total cost of the design, from initial specification to PG tape), which ranges from 5M$ to 50M$ for a completely new complex SoC design using *today's* 0.13 micron technologies. Using the same assumptions as above, this implies volumes of 5 to 50 million chips to break even.

These figures partially explain the strong growth rate of Field-Programmable Gate Arrays (FPGA) and application-specific standard products (ASSP) in certain markets. This is particularly true in the communications infrastructure space for example, where medium volumes (below 100K chips/year) preclude the development of specialized ASIC's, with the exception of high margin systems business, where high performance can command higher prices.

A radical change is needed [1] to allow small-to-medium entrants in the market, or to support products with volumes well below the multi-million chip threshold needed to make a profit on low-cost IC's. Somehow, the mask-set NRE needs to be reduced or amortized over many more products. FPGA's are one solution to this, but their higher power and cost preclude high-volume and low-power applications. Recent approaches using a gate-array style fabric and top metal-level configuration will also help provide an intermediate point on the NRE-flexibility continuum.

However, neither of these solutions address the *design* NRE and time-to-market needs for today's SoC's which can have over 100 million transistors – enough to theoretically place the logic of over one thousand 32bit RISC processors on a single die. Leveraging these capabilities is a major challenge. For this reason, an SoC design platform needs to be amortized over many variants and generations of a product family, to help amortize both the mask *and* the design NRE's. Moreover, platform users need better productivity tools to reduce the end-product design NRE.

In light of the above, the required SoC design paradigm change will be driven by three key requirements:

1. *Faster time-to-market for SoC platform implementation*. In particular, through the use of higher-level off-the-shelf IP's, connected via a

1530-1591/04 $20.00 (c) 2004 IEEE

modular, scaleable SoC interconnect topology and standard communication interfaces.

2. *Increased flexibility in SoC platforms* to amortize the mask and design NRE over more products. This can be achieved by a combination of S/W programmability and (re)configurable H/W, leading to more reusable platforms.
3. *Dramatically increased productivity for the platform user*. This will be the key requirement and will have the highest impact on the application's specification and the underlying platform architecture. This will drive the development of new approaches to facilitate automated application-to-platform mapping.

Emerging SoC platforms will include a wide diversity of IP components. These include software programmable processors - from general-purpose RISC to specialized application-specific instruction-set processors (ASIP); platform FPGA's, combining coarse- and fine-grain reconfigurable logic, memory blocks and reconfigurable I/O; embedded FPGA and/or sea-of-gate fabrics; and finally, traditional semi-custom or custom H/W. Each represent different trade-offs in time-to-market versus product differentiation (power, performance, cost), as depicted in Figure 1 below.

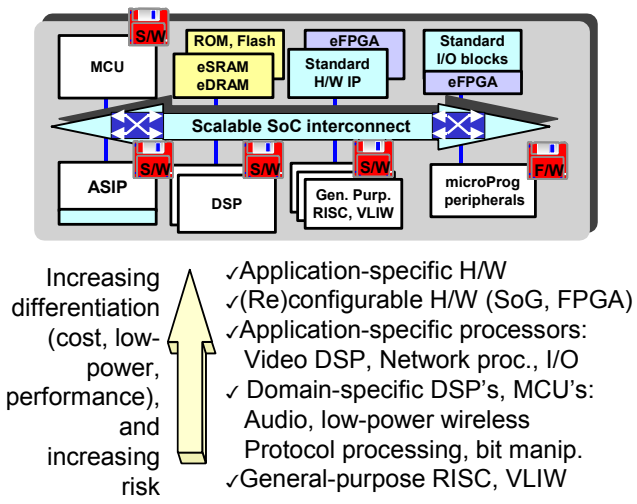

Figure 1. MP-SoC Platform Component Options

This panel addresses a range of approaches to the emerging SoC economic and technical dilemma, from the viewpoints of a range of players:

- a telecom system house,
- a data communications house,
- a consumer-oriented system house,
- a consumer electronics design group of an SoC design and manufacturing company,
- finally, an academic researcher presenting a case study of different leading approaches for a wireless 3G application.

2. From ASIC's to FPGA's: from the perspective of fixed and mobile telecom infrastructure systems

Jean-Michel Balzano

H/W Competence Center,
Fixed Communications Group,
Alcatel
Lannion, France
Email: jean-michel.balzano@alcatel.fr

Alcatel fixed and mobile telecom infrastructure divisions are major users of a complete set of available IC's, ranging from ASSP's, ASIC's to FPGA's. The evolution of the market environment of these divisions, combined with the ICs offering evolution, has lead to an original path and specific requirements. This section shows how this industry has migrated from ASIC's to ASSP's and FPGA-based products, and what could be the next move.

The early 90's telecom landscape was made of equipment providers, designing proprietary equipment usually linked to the specific requirements of the national operator, and eventually selling their products to countries with no telecom industry. Standards were limited to the interoperability of systems from different vendors, and internal interfaces remained proprietary. Consequently the ASSP's offering was limited because of a restricted market. In addition, time to market was not a top priority consideration, due to captive markets for both the operator and the equipment vendor. This situation created an ideal target for ASIC's, as an example more than 30 different ASIC designs are embedded in a core switch developed in the 90's.

Progressively, along with the deregulation of this market, which encouraged competition and open standards, telecom oriented ASSP's started to flourish, such as framers or switches. Along with competition, time to market was becoming more and more critical and the rise of FPGA's has been an adequate answer to this challenge. Because of the comparatively low volume of infrastructure products and despite higher price, FPGA's more than balanced the NRE cost of ASIC's, that is doubled at every silicon geometry reduction step.

In addition, ASIC design turned to be more and more time-consuming and expensive, extensive design verification was extremely long, CAD tools were extremely expensive, all this making ASIC's less attractive. Another factor, linked to the low volume of this sector, was the low interest

from ASIC's vendors for this market, and finding an ASIC was becoming challenging. As a consequence, infrastructure products designed in late 90's are essentially based on ASSP's and glue FPGA's. These IC's are currently fulfilling all the needs, in terms of capacity and speed, of these products leaving no room for ASIC's.

After the major downturn of the overall telecom industry, infrastructure equipment is now facing two challenges. On one hand, keep the current legacy products (i.e. TDM switches) alive and competitive, on the other hand, develop the next generation products (UMTS, NGN…).

Keeping legacy products alive means phase-out resolution including replacement of ASIC's from old technologies. This is a great opportunity for complex FPGA's, that now have the required capacity for this job. This also means cost optimization, for both CAPEX (capital expenses) and OPEX (operating expenses). Solutions optimized, for the 1K to 10K parts per year volume range, for low power dissipation and compactness are needed there.

Next generation infrastructure products are based on the clear separation of the control plane from the user plane, with a very high level of processing either on the voice and the packet domains. This processing is currently done by DSP's or NPU's providing a high level of flexibility, at the price of a comparatively poor capacity, performance over cost ratio.

Bridging the gap between flexibility and high performance, in this area, could be the next challenge for FPGA's evolution and their use in this industry. This is only possible with a good combination between adequate features on the chip (SoC, DSP-like features, etc), and powerful tools providing the same abstraction level as software tools (C coding, co-verification), but tailored to the FPGA context in terms of complexity and price.

3. Perspective of data communications house

Allan Silburt

Carrier Core and Multi-Service Business Unit
Cisco Systems Inc.
Ottawa, Canada

Cisco products cover a wide range of customers from home networks to internet core routers. The business factors that drive silicon technology choices cover a wide spectrum. My perspective emerges from the CCMS (Carrier Core and Multi-Service) Business Unit which covers the high performance core internet end of our business. Technology drivers for us at 90nm are first derived from business drivers – what moves our customers to be profitable.

Market leadership in core routing means serving large service providers and private network builders. Their business is driven by providing high value added services – in addition to raw bandwidth – things like security, private networks and Quality of Service features. Products they install in their networks have fairly long lifetimes and the operational costs tend to heavily outweigh the up front capital costs. The result is a demand for feature rich, but highly programmable architectures that can be incrementally enhanced with new features over time to protect their investments and leverage the huge software base of Cisco's IOS operating system. While we strive to optimize time to market and cost for the silicon we embed in these products, the fundamental viability of a high end product rests on features, power dissipation and programmability. We are committed to make the substantial investments in silicon that distinguish our products, embed our intellectual property and allow us to deliver superior value.

ASSP's typically populate the boundaries of our proprietary engines and data paths at standardized interfaces and physical layers. FPGA's continue to be the glue that holds the functions that fall outside the logical system partition lines. High complexity FPGAs, with their relatively higher cost/power per complexity profile, find their way into point applications driven by time-to-market priorities. These products are typically built entirely from ASSP technologies including some commercial network processors. Platform FPGA's – in compromising the optimal power/density/speed possibilities in favor of some schedule improvements are not likely to have an impact on data paths in our high end systems. However, they are making inroads on the control plane.

Our main pull on 90nm silicon technology is driven by our core engine technologies and their supporting chips. Here we need to exploit power/speed/density opportunities of 90nm with Cisco proprietary processing architectures augmented with 3rd party cores that form the intellectual property of our routing engines and switch fabrics.

As for CAD tool challenges at 90nm, timing analysis capability that accurately models the many complex secondary effects that become significant at this node with reasonable execution times for a large chip is a big concern of ours. Since compute HW capability follows the same growth curve as our chip gate counts (Moore's Law) we are at risk when new effects introduce non-linear computation complexity.

4. Perspective of consumer electronics system house

Kees van Berkel, Research Fellow
Philips Electronics[1,2]

[1] Philips Research Labs, Prof. Holstlaan 4,
5600 AA Eindhoven.
Email: kees.van.berkel@philips.com

[2] Technical University Eindhoven,
Dept. of Computing Science and Mathematics.

Digital signal processors have made a major contribution to the success of GSM handsets: for many critical channel and voice processing functions DSP's offer the required flexibility at competitive cost and power consumption levels. For 3rd generation mobile handsets, however, their performance falls one to two orders of magnitude short for functions such as rake reception and P-SCH searching. Dedicated accelerators, however, lack the flexibility to track the evolution of the standards (e.g. for UMTS release R5, R6) and to enable stepwise improvements of algorithms (e.g. interference cancellation). Furthermore, there are multiple 3G standards, older standards have to be supported (GSM, Edge, GPRS), and future handsets may be equipped with WLAN functionality (802.11a,b,g), broadcast reception (DVB), positioning (GPS), and a range of multi-media functions (JPG, MPEG4, etc).

With these kind of requirements in mind, I will argue that there is a need for a fourth category of computing, positioned between dedicated hardware (possibly weakly configurable) on the one side, and firmware (on a DSP) plus software (on a micro-controller) on the other side. This fourth category, let's call it *crunchy ware*,

- must approach the flexibility of software (in order to meet the above requirements),
- must approach the cost and power levels of dedicated hardware,
- for a given application domain (in this case mobile communication).

Future mobile handsets will contain all four categories: hardware, crunchy ware, firmware, and software. The amount of code for each category [bytes] could scale as 1kB .. 30 kB .. 1MB .. 30MB. Likewise, processing power could scale as 100 GOPS .. 10 GOPS .. 1 GOPS .. 100 MOPS. With each technology step, functions will have a tendency to migrate towards a more flexible category.

A key notion is *flexibility*. Webster's defines *flexibility* as "characterized by a ready capability to adapt to new, different, or changing requirements". This definition indeed quite well reflects the requirements at hand. A key challenge is, however, how to characterize (and preferably quantify) this flexibility, both as *required* by the different applications, and as *offered* by crunchy ware. Major themes here include: task granularity, memory access patterns, data parallelism, and branching.

I'll argue that reprogrammable/reconfigurable vector processing (SIMD) is a major candidate for crunchy processing. For a vast range of signal-processing kernels it offers abundant data parallelism, and high memory bandwidth at a remarkable flexibility and efficiency (cost and power).

Accordingly, a main challenge for EDA is vectorizing compilation (e.g. from C/MatLab, with special attention for reduction operators, vector shuffles, and scalar code).

5. Perspective of consumer electronics design team of SoC design and manufacturing company

Richard Bramley, R&D director,

DVD Division,
STMicroelectronics
Email: richard.bramley@st.com

With rising NRE costs and complexities in sub-micron technologies few would argue with the need to do more with less just what is meant by more and what is meant by less.

My standpoint in this area is firmly founded in direct experience from the cutthroat world of consumer electronics. The feature expectations and price-pressure in the DVD player market over the last few years are largely unprecedented in the industry, and as such, can serve as a rich example.

Over the last 2 years the kit price for the silicon in a DVD player has declined over 60%. In the same timeframe, many new audio and video algorithms have become standard in all players (Divx, WMA, MP3, etc.). Many of these new features are entering the consumer world directly from the PC, which, for easy implementation, dictates a largely processor-based strategy. Audio decoders are moving to longer processing windows and are taking advantage of the 32-bit world of the PC where they are developed. Indeed, the balance of processing power and complexity is no longer in the decoders. It now resides in the post-processing algorithms. Video decoding is becoming extremely complex with the multiplication of prediction modes, fractional pixel processing and increased conditional processing load.

With this as a backdrop I will argue for the use of standard (as opposed to media) processors in these devices, the specialized processing for audio & video being implemented using specialized accelerators, themselves programmable (e.g. audio sample rate conversion, predictor filtering). As specialized hardware is replaced by standard processors the devices we make become more homogeneous and therefore, more manageable in terms of design complexity. The programming environment also becomes more uniform and homogeneous.

Today's processor designs allow us to do this at no extra cost in silicon area, but also bring a whole set of new EDA problems to solve. Particularly in the areas of system load modeling and hardware/software partitioning when system software is not complete.

6. A 3G Wireless Application Case Study

Norbert Wehn

Microelectronic System Design Research Group
University of Kaiserslautern
Erwin-Schrödinger-Straße
67663 Kaiserslautern, Germany
Email: wehn@eit.uni-kl.de

In wireless digital communications, bandwidth and transmission power are critical resources. Therefore, advanced communications systems have to rely on sophisticated forward error correction schemes which are key building blocks in the outer modem of baseband signal processing. Turbo-Codes belong to the most efficient error correction schemes and are part of existing and emerging communication standards (e.g. UMTS). Requirements on different quality of services, multi-mode support, and varying throughput, demand flexibility in terms of programmability or reconfigurability. On the other hand, terminal applications, which are driven by cost and mobility, require minimum energy and footprint which can only be met by dedicated application-specific solutions.

We have investigated the design space of Turbo-decoders for different implementation styles with varying throughput requirements and compare them with respect to architectural efficiency and corresponding design effort. All investigations are based on a technology under worst-case conditions, 3GPP decoding conditions and Log-MAP based Turbo-decoding.

6.1 Configurable and DSP Cores

In [2], we have exploited the capabilities of the application-customized RISC core Xtensa [3] to obtain high-performance implementations of the Turbo-Decoder algorithm. The algorithmic bottlenecks were removed by adding application-specific commands (double butterfly operation, zero-overhead data transfers) to the RISC cores, which increases the instruction-level parallelism. The Xtensa core, without instruction set extensions, requires 48k gates and runs at 180 MHz. The core with the application specific instructions needs 104k gates running with 133 MHz.

Table 1 summarizes performance results in kbit/s for different optimized implementations, in order of increasing application-specific functionality:

- the ST120 general-purpose DSP of STMicroelectronics,
- the TigerSharc from Analog Devices, which has a specific instruction support for the MAP algorithm, and
- the Xtensa core with the instruction-set extensions described above.

Table 1: Comparison of mono-processor solutions

Processor	Features	Clock [MHz]	Throughput [kbit/sec]
ST120	VLIW, 2 ALU	200	200
ADI TS	VLIW, 2 ALU (w. MAP support)	180	666
Xtensa	Config. RISC (w. a-s instructions)	133	1400

The total design effort for the Xtensa core implementation adds up to about 4 man-weeks for the concept phase, derivation of a bit-true C-model and system validation, and 3 weeks for the instruction enhancements, their validation and implementation.

6.2 Application-Specific Hardware

An application-specific hardware with the same degree of parallelism yields about 5 Mbit/s throughput and consumes about 25Kgates. The design effort for this architecture is about 4 man-weeks for concept phase, bit-true C-model and system validation, and 3 man-months for VHDL coding, synthesis and validation for a skilled VHDL designer.

6.3 Application-Specific Parallel Hardware

Applications which demand higher throughput require massively parallel architectures, independent of the implementation style. In [4], we presented a parameterizable and scalable turbo-decoder architecture. Parallelizing the interleaver network is key for this high throughput turbo-decoder architecture. The decoder was implemented as a fully parameterizable VHDL model (parameters are: type of decoding algorithm, window and

acquisition length, maximum blocklength, parallelization type and degree). Table 2 presents results for various architecture derivatives, running with a frequency of 166 MHz using the Synopsys Design Compiler.

Table 2: Comparison of dedicated H/W configurations

Parallel MAP Units Parallel I/O	1 1	4 1	6 2	8 2
Total area [mm^2]	3.9	9.2	13.0	17.3
Energy per block [µJ]	48.7	51.7	50.9	55.2
Throughput [Mbit/s]	11.7	39.0	59.6	72.7
Efficiency (norm.)	1.00	1.32	1.47	1.24

The design effort for this architecture adds up to 4 man-months for concept phase, bit-true C-model, and 9 man-months for VHDL coding, synthesis and validation for a skilled VHDL designer.

6.4 Application-Specific Multi-processor

In [5], we investigated the implementation of an application-specific multiprocessor solution based on the Xtensa core discussed in the previous section. The core was extended by a specific single-cycle interface for fast inter-processor communication. A heterogeneous communication network was developed which is based on a similar structure as used in the scalable application-specific architecture. Table 3 presents results for different number of Xtensa processors.

Table 3: Comparison of multi-processor configurations

Number of Processors	Throughput [Mbit/sec]	Area [mm^2]	Efficiency
1	1.48	6.42	1.00 (0.08)
8	11.58	20.91	2.58 (0.18)
16	22.64	36.98	2.66 (0.20)
32	43.25	70.26	2.67 (0.21)
40	52.83	87.47	2.62 (0.20)

The efficiency in this table does not take into account the energy since we had no access to power characterization of the Xtensa core. The efficiency value in brackets is normalized to the application-specific dedicated H/W solution. The design effort for this multiprocessor implementation (assuming that we started from scratch) adds up to 4 man-months for the concept phase, the C-model and system validation, 1 man-month for the adaption and programming of the Xtensa core and 3 man-months in total for the application-specific communication network.

6.5 FPGA Implementation

We also implemented one instance of the scalable ASIC architecture on a Xilinx Virtex II-3000 architecture. The VHDL code developed above was adapted and optimized with respect to the FPGA architecture. The additional design effort added up to 2 man-weeks for memory adaptation, and 1 man-month for VHDL recoding. We used a parallelization degree of 4 parallel MAP units. The FPGA utilization results are as follows:

- 83% of the slices,
- 37% of the FFs,
- 70% of the LUTs, and
- 41% of the block RAMs are used.

The circuit corresponds to approximately 3.1 million gate equivalents. It runs at 88,2 MHz, which yields a throughput of 22 Mbit/s. The same throughput can be achieved by the dedicated H/W solution with a parallelism degree of 2 MAP units.

7. Panel Conclusion

The continued increase in the non-recurring expenses for the manufacturing and design of nanoscale systems-on-chip (SoC) is leading to the need for significant changes to their design and manufacturing. A range of approaches are presented in this panel, presenting the viewpoint of industrial telecom, datacom and consumer system houses, a semiconductor design and manufacturing company, and an academic researcher.

8. REFERENCES

[1] P. Magarshack, P. G. Paulin, "System-on-chip Beyond the Nanometer Wall", *Proc. of 40th Design Automation Conference (DAC)*, Anaheim, June 2003.

[2] H. Michel, A. Worm, M. Muench, N. Wehn, "Hardware/Software Tradeoffs for Advanced 3G Channel Coding", *Proc. of Design Automation and Test in Europe (DATE)*, Paris, March 2002.

[3] Tensilica Inc. http://www.tensilica.com

[4] M. J. Thul, F. Gilbert, T. Vogt, G. Kreiselmaier, N. Wehn, "A Scalable System Architecture for High-Throughput Turbo-Decoders", *Proc. 2002 Workshop on Signal Processing Systems (SiPS '02)*, pp. 152-158, San Diego, Oct. 2002.

[5] F. Gilbert, M. Thul, N. Wehn, "Communication Centric Architectures for Turbo-Decoding on Embedded Multi-processors", *Proc. of Design Automation and Test in Europe (DATE)*, pp. 356-361, Munich, March 2003.