Prospects for Reconfigurable Systems

In 2003, the authors predicted that reconfigurable systems would emerge within 10 years. They expected the arrival of a "Holy Grail" memory component and chip and wafer stacking. These advances did not occur. Reconfigurable systems should appear in markets where equipment has access to continuous power, but financial incentives and the lack of a versatile memory component inhibit their emergence in mobile systems.

Nick Tredennick
Brion Shimamoto

A little over 10 years ago, we wrote "The Inevitability of Reconfigurable Systems" for the October 2003 issue of *ACM Queue*.[1] We expected reconfigurable systems to be widespread within 10 years. This didn't happen. It didn't even happen in mobile devices, where we saw the greatest technical need. In this article, we review our predictions and reassess the prospects for reconfigurable systems.

Back in 2003

Here, we review predictions we made in the following areas: the value PC market, memory chips, programmable logic devices, the value transistor market, wafer stacking, microprocessors and digital signal processors (DSPs), and design tools.

The value PC market

Prediction: The personal computer attains "good-enough" status. After more than 20 years of semiconductor advances, the PC's performance satisfies most users. Users now shop for value, not for performance, thus lowering profit margins in the PC business. With lower margins, hardware manufacturers shift their design emphasis and their investments to an emerging market: mobile devices. The shift from tethered (continuously powered) systems to mobile devices shifts the design emphasis from price-performance to price-performance per watt.

For decades, the PC dominated the semiconductor business. PCs, at one time, consumed 40 percent of worldwide semiconductor production. PCs began as niche, business-oriented products. New PCs sold because they were faster. This demand for performance drove the PC market to more capability and higher clock frequencies. PC performance grew at an exponential rate, following the Moore's law rate of improvement in semiconductor components.

The demand for performance remained strong for more than 20 years, but the customer set increasingly included users who were satisfied with less than leading-edge performance. The market grew, and this gap in the demand for performance, represented by users at the leading edge and by users at the trailing edge, widened. How did it play out? The PC's performance ("Supply" in Figure 1) grew much faster than the bulk of performance demand, leading to the "good enough" PC.

Published by the IEEE Computer Society

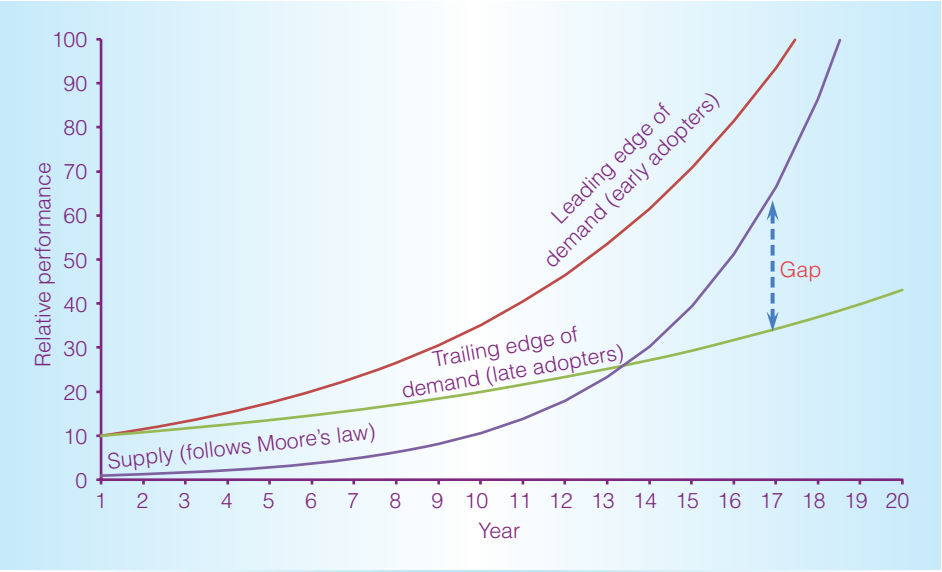

Figure 1. Leading-edge and trailing-edge demand spread. At the PC's introduction, it was slow and not very capable. Over time, the PC's performance increased at a rate related to Moore's law improvements in its underlying semiconductor components. Simultaneously, the user community's demands spread between the rapidly growing demands of the PC's early adopters (at the leading edge of applications) and the slowly growing demands of late adopters (applications such as editing, inventory tracking, and bookkeeping). Demand spread also because PCs at the leading edge tackled ever-harder, computationally intensive problems while cheaper PCs at the trailing edge served a wider range of simple applications. Some years after the PC's introduction, its increasing performance exceeded the performance demanded by users and applications at the trailing edge.

Memory chips

Traditional memory components—static RAM (SRAM), DRAM, and flash—are unsuitable for mobile devices. SRAM is fast, but it's expensive and power-hungry. SRAM and DRAM retain their content only as long as they have power. DRAM is dense, but it's slower than SRAM. Flash memory retains its content even after a power loss, but flash is slow to read, very slow to write, and it wears out. The PC's design leveraged the advantages of each memory type while masking its weaknesses. But doing that isn't possible with mobile devices, because the power dissipation in the hierarchy of components itself is a problem.

Prediction: Because of the mismatch between power-sensitive mobile devices and available memory components, the "Holy Grail" memory component will arrive in "two or three years."[1] This component will have the speed of SRAM, the density of DRAM, and the nonvolatility of flash memory.

In 2003, the candidates for the Holy Grail memory component included ovonic unified memory (a phase-change memory), ferroelectric memory (FRAM), magnetoresistive memory (MRAM), memristor-based resistive memory (RRAM), carbon nanotube memory, and molecular memory. Major semiconductor companies such as IBM, Intel, Motorola, Samsung, STMicroelectronics, and Texas Instruments divided into camps and were investing heavily. Numerous start-ups pursued the Grail.

Programmable logic devices

Prediction: Manufacturers of programmable logic devices (commonly referred to as field-programmable gate arrays [FPGAs]) will benefit greatly from Holy Grail memory. It enables denser, cheaper FPGAs with nonvolatile configuration memories. Nonvolatile

configuration memories are important for system security (because there's no external access to the configuration bits), for reliability (because there are fewer components), and for speed (because configuration bits don't have to be loaded during system initialization). FPGAs are a critical component in reconfigurable systems. Such great improvements in FPGAs will signal major growth in the market for reconfigurable systems.

The value transistor market

Prediction: Semiconductor advances tracked Moore's law for decades. Semiconductor customers are similar to PC customers. All semiconductor applications initially demand more performance. Later, transistors will improve with Moore's law and the market will expand to include less-demanding applications. For a growing range of applications, transistors will improve faster than the demand for their performance rises. Eventually, the *value transistor*—a transistor that's good enough for most applications—will emerge. The value transistor favors foundries over integrated device manufacturers because foundries amortize their production over a larger base of applications. If most applications use value transistors, then semiconductor foundries will have become good enough. The market for state-of-the-art semiconductor processing equipment will cease to be a leading indicator of industry health.

Wafer stacking

Prediction: Semiconductor manufacturers will move their investment from transistor shrinking to 3D circuits—in particular, to wafer stacking. (With chip stacking, the stacking costs are paid per chip, while wafer stacking amortizes the stacking costs over the number of chips on the wafer, giving wafer stacking more favorable economics.) Wafer stacking is the most promising avenue for integrated-circuit performance improvement.

Going vertical has many advantages. The foremost advantage is shorter interconnects, which mean smaller drivers and receivers and fewer problems with metal migration. (Vertical interconnections can have large cross-sectional areas.) Smaller drivers and receivers use less energy; shorter distances mean delays fall and processing speed rises.

In addition, it's possible to optimize the semiconductor process to the application. Build one chip in a pure logic process, another in a pure memory process, and a third in a pure analog process, and stack them to build the system. There are no compromises forced by mixing processes.

The next big advantage of going vertical is the reduced physical footprint, which is of primary importance in mobile devices. Think of Manhattan's footprint and commute times if no apartment or office stacking was allowed. The area increase would be enormous; commute times would increase exponentially (exactly what hampers communication links in 2D chips today).

With three dimensions, it's possible to build structures, such as inductors. Inductors are needed to make oscillators, and oscillators are used everywhere. Inductors cannot be built in the X-Y world of planar semiconductors.

With wafer stacking, FPGA manufacturers would get denser, lower-power FPGAs.

Microprocessors and DSPs

Prediction: Microprocessors and DSPs are unsuitable for cost-performance-per-watt applications.

There are three categories of hardware: custom hardware, microprocessor-based hardware, and reconfigurable hardware.

Custom hardware solves a particular problem directly. That is, the algorithm that solves the problem is embedded in fixed, special-purpose hardware resources: fixed, special-purpose hardware with fixed algorithms.

Microprocessor-based hardware solves a range of problems because it provides fixed, but general-purpose, resources that can host a range of algorithms. Memory stores programs written in the microprocessor's instruction set: variable algorithms running on fixed, general-purpose hardware. The great breakthrough of the invention of the computer was in separating the algorithm from the hardware. This separation enabled the creation of standard components (processor, memory, peripherals) that supported a range of applications. The separation raised the level of abstraction in design (to programming)—an act that improved engineering productivity.

Reconfigurable hardware solves problems by enabling changes in the hardware and in the algorithms: variable, special-purpose hardware with variable algorithms.

The reason that microprocessors (and DSPs) are unsuitable for mobile applications is that they are inherently inefficient. The microprocessor's instructions simulate only the behavior of a desired system; custom hardware is the system in circuits. The microprocessor is still necessary in mobile devices to manage what is going on. But computationally intensive tasks belong in custom hardware for efficiency's sake.

Design tools

Prediction: New FPGA software will enable software engineers to do circuit-based design. The main thing holding back reconfigurable systems is that their use requires logic design expertise. Design tools to translate behavioral descriptions into reconfigurable systems would enable software engineers to design reconfigurable systems. Software engineers outnumber logic designers by at least 10 to one.

2003 summary

The change in emphasis from PCs to mobile devices changed the design goal from cost-performance to cost-performance per watt. The power-sensitive design goal rendered traditional microprocessors and memory components unsuitable for mobile devices. Mobile device makers are making do with today's components; that's all they have.

Increased investment in memory components should result in a Holy Grail memory component. This would greatly improve both FPGAs and the cost-performance per watt of mobile devices. Investment in the semiconductor process for 3D integrated circuits would improve both memory and FPGAs. Holy Grail memory and FPGAs with nonvolatile configuration memory would mean substantial growth for reconfigurable systems.

What we got right (or what was obvious)

PCs are good enough; the value PC's emergence shifted the design emphasis to mobile devices. The design goal for mobile devices is cost-performance per watt. The memory hierarchy of SRAM, DRAM, and flash memory is unsuitable for mobile devices, and microprocessors are unsuitable for mobile devices. That's about it.

What we got wrong

The Holy Grail memory component is not here. The 3D integrated circuit is in its infancy, and wafer stacking is not here. There are some reconfigurable systems, but they aren't pervasive, and they have made no inroads into mobile devices.

Status

Ten years after our prediction, mobile devices are still implemented with SRAM, DRAM, and flash memory. They still use inefficient microprocessors and DSPs even for computationally intensive work. The breakthrough memory component has not occurred. Without a Holy Grail memory cell, the FPGA can't improve sufficiently to enable mobile applications.

Nonvolatile memory

The Holy Grail memory cell faces formidable challenges. Read and write speeds remain slow. Some memory candidates have limited persistence or insufficient lifetimes. Some use too much power, and some suffer consistency problems over the memory's life cycle. Most of the candidates use nonstandard semiconductor processes, which handicaps them with high, uncompetitive costs. Finally, all of the nonvolatile memory candidates lag far behind the incumbents' learning curve. That is, the gaps in cost, performance, and capacity between any Holy Grail memory component and traditional SRAM, DRAM, and flash memory are too great to foster displacement by the new memory component. Despite the enormous financial incentive awaiting the Holy Grail memory cell, the industry seems little closer to the goal today than it was in 2003.

Wafer stacking

Despite financial incentives on par with those for Holy Grail memory, there are few examples of progress in wafer stacking since 2003. Micron's Hybrid Memory Cube

(http://www.micron.com/products/hybrid-memory-cube) is one example of chip stacking. The list of challenges to the widespread adoption of wafer stacking is even longer than the list for nonvolatile memory cell development.

- Wafer bonding, such as low-temperature metal-oxide bonding, joins chips on separate wafers together. It has to bond reliably and uniformly, and must bond without requiring heat that could degrade the chips' circuitry.
- Precision alignment is critical for matching the vertical connections between chips. As the through-silicon vias (TSVs) get smaller, their alignment must be correspondingly precise.
- Surface conditioning is an issue: wafers must be perfectly flat for the chips to bond properly. Surface properties must be precisely controlled.
- Vertical etching drills the holes that become the TSVs. In the best case, it would be possible to etch deep, small-diameter holes. However, stacked wafers are being thinned to a few microns because of the trade-off between wafer thickness and TSV pitch; thicker wafers imply larger TSV pitch (fewer interconnects per unit area).
- Thermal management is an obvious issue. Stacking 16 or 32 chips into a volume and surface area that was formerly occupied by one chip package invites cooling problems. Differing coefficients of thermal expansion stress the connection points at chip interfaces. The 3D-chip packages are no taller than 2D-chip packages!
- Design tools don't support 3D design well.
- In terms of standards, if the industry is to develop open sources for building stacks from chips supplied by diverse manufacturers, there must be standards that facilitate cross-company product flow to the stack integrator.
- Stack yield is less than the product of yields for the individual layers. Yield needs architectural support in the form of redundancy, real-time configuration, self-checking, and fault tolerance. In a stack of 32 chips, a yield of 98 percent on individual chips drops the stack yield below 50 percent.
- Testing challenges increase for stacked chips.

In addition to the technical challenges, formidable cultural challenges inhibit 3D-chip development. Within the circuit- and logic-design communities, issues with chips and wafers are often considered to be "packaging" and to be in the province of line manufacturing. Organizations that have traditionally been mostly isolated from each other (design and manufacturing) will have to mix.

Summary

Our 2003 forecast was humbled by our big misses: the Holy Grail memory cell and wafer stacking. Without these two advances, FPGAs have been unable to improve sufficiently to invade mobile applications. This has kept opportunities for reconfigurable systems in the high-performance computing realm, where systems access continuous power.

What's holding reconfigurable systems back

The question "what's holding reconfigurable systems back?" has two answers, one for mobile devices and one for tethered systems.

Mobile devices

In 2003, we outlined a case for compelling demand. We overlooked a tenet of economic thought: demand doesn't create supply. After delineating the case for demand, we should see if suppliers have enough incentive to fill the need. The suppliers, in this case, would be the FPGA companies, primarily Altera and Xilinx. They own the rights to the FPGA fabric that is the critical element of reconfigurable systems. These FPGA companies could build configurable chips for mobile devices or they could license the FPGA fabric to mobile-device chip suppliers.

The growth in mobile devices places increasing demands on the mobile

communications infrastructure: base stations, communications radio links, switches, and routers. These tethered communication systems have been the core market for FPGA companies for decades. This market has been growing rapidly. It demands high-end components, which mean high margins. The FPGA companies, therefore, concentrate on those high-performance, high-capacity, communications-oriented components that deliver excellent profits.

The mobile-device market, by contrast, is a high-volume consumer market, which implies highly competitive, low-power designs and thin margins. So far, the FPGA companies have not licensed their FPGA fabric to the mobile-device market's suppliers. Major FPGA companies have had neither the inclination nor the spare engineering talent to build chips for mobile applications. The major FPGA companies are unlikely to venture into reconfigurable mobile devices.

This might seem like an opportunity for a start-up, but there's still the crippling lack of Holy Grail memory. We see no near-term prospect for Holy Grail memory. That aside, prospects are still dim for a breakthrough start-up in reconfigurable systems for mobile devices. The reason is that, while the FPGA business seems like a component business that would be amenable to a better mousetrap, it isn't. The FPGA business is not really an FPGA-component business; it's an FPGA-software business. The major FPGA companies have invested engineering millennia developing the software support environment for their FPGAs. No start-up has the time, money, or engineering talent to create competitive FPGA support software. Even a large company couldn't do it on its own.

For the time being, mobile devices will have to muddle along without reconfigurable application support from Altera or Xilinx.

Tethered systems

In tethered systems, reconfiguration is emerging.

Components from the FPGA companies began as "glue logic": FPGA fabric that could sweep up the miscellaneous logic on a circuit board. FPGA chip capacity grew enough to hold substantial logic subsystems, but they were still primarily programmable logic, though with some hard macros such as multipliers and blocks of memory to make logic designs faster and more efficient. Over time, the hard macros—processors, transceivers, and memory blocks—became more important, more numerous, more capable, and more standardized. In the last few years, there has been a subtle but significant transition: the FPGA is no longer a component with hard macros; it is a system on a chip (SoC) with FPGA fabric.

These FPGA-based SoCs are increasingly accessible to software engineers because they look like microprocessor-based SoCs, with the addition of FPGA fabric. Hardware and software engineers are familiar with building systems and with programming applications for microprocessor-based SoCs. That means the major FPGA makers are breaking away from their limited base of hardware logic designers and are on a path to enable the much larger base of software engineers.

Specialization and intellectual property (IP) reuse portend success as well. Some engineering firms specialize in creating and licensing efficient computationally intensive IP macros for the FPGA fabric. FPGA companies have an inherent advantage no one else has. With today's transistor densities, FPGA makers can integrate anyone's logic on chip next to their FPGA fabric. "We can do everything you can plus…"

Enabling software engineers for FPGA-based design isn't risk-free; it requires a bridge between the FPGA and the software engineer and will demand something of the software engineer. That bridge may be the Open Computing Language (OpenCL), a C/C++ language that facilitates mapping applications onto multiprocessors, GPUs, and FPGA fabric. OpenCL is a good candidate for enabling software engineers to build reconfigurable tethered systems.

Our original thinking saw a progression from the logic designer's custom hardware (fixed hardware with fixed algorithms) to the programmer's microprocessor-based system (fixed hardware with variable algorithms). From there, we saw a new design environment that required hardware engineers to split the application between programming segments (variable algorithms)

and performance-critical segments residing in the FPGA fabric (variable hardware). We thought the design of reconfigurable systems required a hardware engineer's understanding of the system architecture and the problem's requirements. The flaw in this thinking is that it required a step backward in abstraction; however, no step backward in abstraction can succeed, because engineering talent and engineering productivity are critical resources. Engineers are more productive at a higher level of abstraction (in this case, programming). Any venture that proposes increasing spending on a critical resource will fail.

Wafer stacking now seems too far out to forecast, but chip stacking is advancing. Current and near-term products aren't quite 3D integrated circuits, but what might be called 2.5D stacks. The inside of what looks like an ordinary integrated-circuit package begins to resemble a miniature circuit board. In the first generation, a passive silicon interposer plays the role of the circuit board. Individual bare-die chips—such as FPGAs, CPUs, application-specific integrated circuits (ASICs), application-specific standard products (ASSPs), analog-to-digital converters (ADCs) and digital-to-analog converters (DACs), and memories—attach to the interposer's top surface. It's similar to the familiar multichip module, but the interposer's TSVs enable much richer interconnections among the pins and chips. With the 2.5D stack, semiconductor manufacturers tailor the process for each chip to its function. Performance-robbing process compromises that are a part of mixing logic and memory processes or of mixing analog and digital functions become unnecessary. So, chips are faster, smaller, and more efficient—they're more suitable for every application. Suppliers can mix and match chips in the package to create a large variety of products without the delays of chip design and verification.

True 3D-chip stacking should follow mature 2.5D stacking. Many of the challenges faced by wafer stacking apply to chip stacking, but the advantages of dense vertical interconnects (lower power, smaller drivers and receivers, faster communication, and shorter wire lengths) provide sufficient incentive for development.

Reconfigurable systems are emerging in tethered applications, notably, in the infrastructure of mobile communications. The mobile device market lacks a large supplier willing to accept the low margins of a consumer market. Don't wait for Altera and Xilinx—the dominant FPGA makers—they enjoy 70 percent gross margins in tethered markets. Tethered markets currently saturate the capacity of Altera and Xilinx, which leaves the mobile-device market to smaller FPGA companies such as Actel, Lattice Semiconductor, MicroSemi, and Tabula. Lattice Semiconductor is expanding its mobile-device applications base, so reconfigurable systems could emerge even in low-margin mobile devices. MICRO

Reference

1. N. Tredennick and B. Shimamoto, "The Inevitability of Reconfigurable Systems," *ACM Queue,* vol. 1, no. 7, 2003, doi:10.1145/957717.957767.

Nick Tredennick is a consultant. His research interests include industry trends, microprocessors, and FPGAs. Tredennick has a PhD in electrical engineering from the University of Texas at Austin. He is a life fellow of IEEE.

Brion Shimamoto is a small business owner. His research interests include microprocessors, computing systems, and design methods. Shimamoto has a bachelor's degree in mathematics from the University of California, Los Angeles. He is a member of the ACM.

Direct questions and comments about this article to Nick Tredennick, 1625 Sunset Ridge Road, Los Gatos, CA 95033; bozo@computer.org.