Microprocessors, Microcontrollers, and Systems in the New Millennium

NEXT-GENERATION ARCHITECTURES MUST BECOME MORE FLEXIBLE TO MEET THE INEVITABLE CHANGES TO MARKET AND CUSTOMER REQUIREMENTS. TO ENSURE FLEXIBILITY, ARCHITECTS AND DESIGNERS NEED TO FACE THESE REALITIES.

Chris Herring
National Semiconductor Corporation

Motivated by necessity, change is inevitable. The buggy whip manufacturer realized this on seeing a Ford Model A drive past. Until the late 18th century, circumnavigation of the globe was a near impossibility without an accurate way to measure longitude. It wasn't until John Harrison dedicated most of his life to developing an accurate timepiece that the problem was solved.

Is fundamental change also a necessity for the microprocessor? Or is there a need to require fundamental change? Where are microprocessors, microcontrollers, and computer systems headed in the new millennium?

Moore's law and its impact

In 1965 when preparing a talk, Gordon Moore noticed that up to that time microchip capacity had seemed to double each year. The pace of change having slowed down a bit over the last few years, we've seen the definition of Moore's law change (with Moore's approval) to reflect that the doubling occurs only every 18 months. Stated another way, transistor count grows at a compound annual rate of 60%.

Process and architectural changes act to drive Moore's law for transistor count and also for performance. CPU cycle time measured in megahertz tends to track Moore's law, following the same path as transistor count. However, system performance is affected not only by a CPU's MHz but also by first memory access time. The CPU can only execute code that it can fetch. This drives the requirement for larger level-1 caches and on-chip level-2 caches found in current CPU architectures.

Figure 1 (next page) shows that the number of transistors used on Intel CPUs has grown 23-fold from 1.2 million on the 486DX2 processor in 1992 to 28 million on the Intel Pentium III with an onboard L2 cache introduced in October 1999. Correspondingly, CPU MHz has increased 20-fold from 50 MHz to 1 GHz during the same time period. Visit http://www.intel.com/pressroom/kits/processors/quickrefyr.htm.

As a check to Moore's law, the expected tran-

0272-1732/00/$10.00 © 2000 IEEE

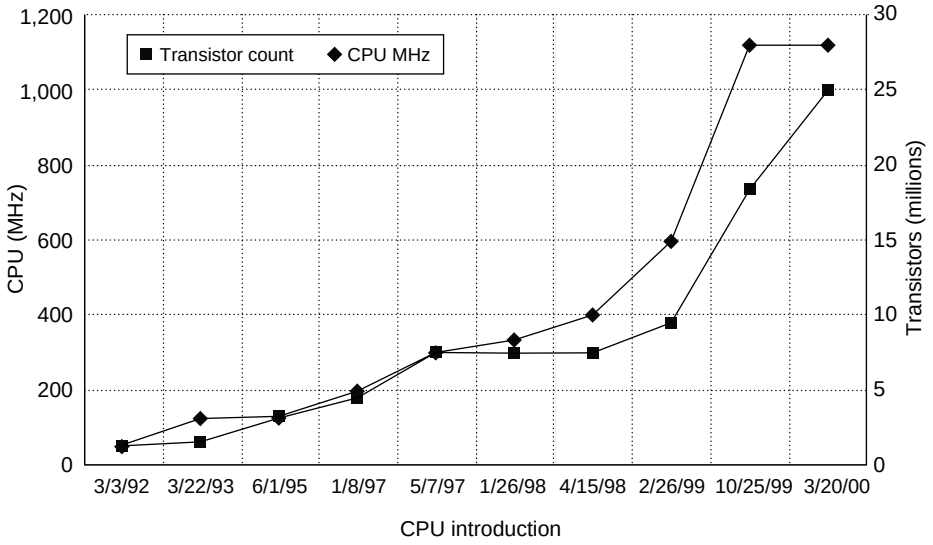

Figure 1. Moore's law showing transistor count and CPU MHz.

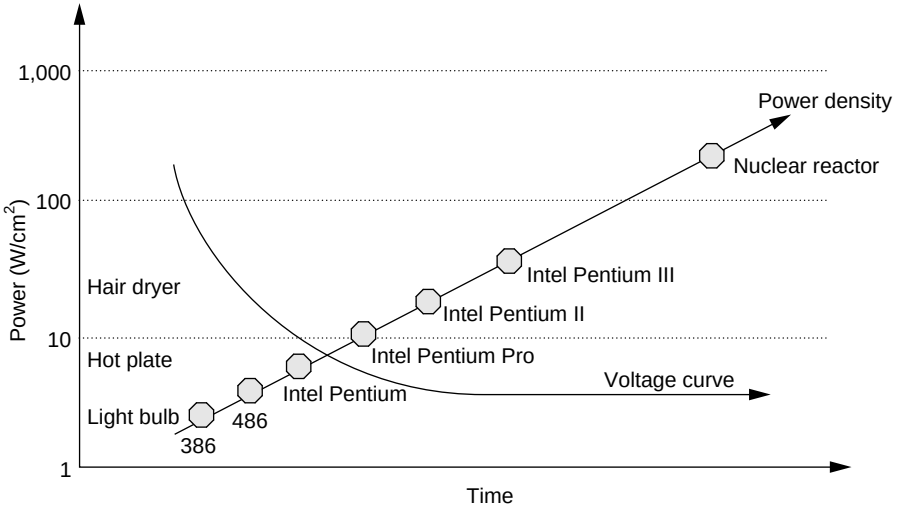

Figure 2. Power density curve.

sistor count starting from the 486DX2 and projected forward 7 years to the Intel Pentium III is 1.2 million × 1.6^7, or approximately 32 million transistors. This is within 12% of the actual figures. Project this forward 10 more years to 2009, and the expected transistor count will approach 28 million × 1.6^{10}, or more than 3 billion transistors with a corresponding expected CPU speed of over 100 GHz!

At the same time, current architectures are becoming more system-on-chip, or SOC, centric. This change to designs incorporating a whole system on one chip implies larger numbers of I/Os will need to escape from the die. If the die size stays relatively constant for a pad-limited design, the available transistor count will explode. Today's 0.18-micron process technology supports a density of 6 million transistors per square centimeter. Current pad-limited designs require approximately 5 to 6 cm^2. Ten years from now the pad-limited design will still be around 5 cm^2, but the transistor density will be 6 million × 1.6^{10}, or 660 million transistors/ cm^2. The pad-limited die will then support 3.3- to 4-billion transistor designs. This leaves quite a lot of free real estate for other functions on the die.

Potential roadblocks

Hard realities set in when considering multibillion-transistor designs. Each new process generation requires finer, more precise lithography techniques, additional and more intricate mask sets, cleaner fabs, and advances in tester capabilities. These designs will be extremely dependent on interconnects on the die. This drives the need for better simulation, modeling, timing, and layout tools. These issues require major design innovations in differing fields, and they must all be available at the same time, or new process advances won't make it to market.

As transistor sizes approach the molecular level, physical boundaries will emerge that may derail the process migration path. Voltage tends to drop with each process improvement, which permits more transistors switching at higher frequencies without a huge increase in power. However, with each new process step, the percentage of voltage drop from one generation to the next decreases.

At some point voltage drops won't offset the transistor growth and switching factor in the power equation (power = capacitance × volt-

age^2 × frequency). Given a fixed voltage, power requirements will begin to grow linearly with transistor count. The overall die capacitance with interconnects, parasitic capacitance, and capacitive coupling will tend to negate reduction in capacitance due to the process technology.

With a doubling in power every 18 months, the costs associated with packaging and thermal dissipation will tend to dwarf any savings achieved with higher transistor densities. In fact, as shown in Figure 2, the power density measured in watts/cm^2 will increase from around 20 W/cm^2 to nuclear reactor densities (250 W/cm^2) as power doubles every 18 months. At this point there's no possible way to dissipate the heat. Additionally, as frequencies increase, noise and coupling issues become more and more difficult to solve, especially with the decreased noise floor immunity offered by lower voltage processes. Finally, market dynamics drive the need for faster design-to-production schedules. (Current power density data can be found from a number of sources; one is http://developer.intel.com/technology/itj/q32000/articles/art_1.htm. The extrapolation to the future is my own doing.)

The world is moving at "Internet speed," yet technology's half-life is measured in months. Today's architecture trend will eventually hit either a brick wall or a number of stumbling blocks. CPU architects can't continue designing by looking in the rear view mirror; they must realize the necessity of change. The convergence of technical and market requirements will act as the catalyst needed to drive invention and the adoption of new CPU and system architectures.

Market dynamics

Market dynamics and usage models are useful considerations when predicting the possible architecture paths for CPUs and systems. Currently, the driving requirement for CPUs is to provide the absolute highest general-purpose performance possible. The performance treadmill of the PC market drives this need. The PC is at its core a general-purpose machine and must fit the needs of a wide range of users from the businessperson to the third-person-shooter 3D game player. The PC's nonapplication-specific nature requires the concept of performance headroom.

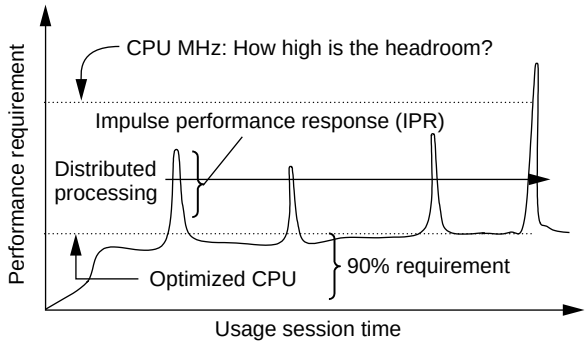

Figure 3. Performance usage model.

Since system architects don't know a priori what the end user applications require, they design a PC with enough performance to accommodate a wide range of usage models. With the deregulation of the Internet in the mid-1990s, the market defined a new usage model in which raw performance began to take a back seat to communication and connectivity. Browsers, plug-ins, and reliable connections are the important performance measures for an Internet appliance—not frames per second or benchmark measurements. The Internet presents a defined usage model with diverging requirements from the PC market, which lacks a defined model.

The usage model for consumer devices indicates that performance is more perceived than measured. Applications require a fixed baseline amount of performance. This equates to typical browsing, word processing, reading, or waiting for input. Interspersed with this is the occasional requirement (an impulse performance response, or IPR) for a short burst of performance needed to play a multimedia file or decompress an image. The old paradigm of designing a system with enough headroom to handle the worst-case performance requirement needs refreshing. The new paradigm lets designers treat the occasional high-performance requirement as an IPR and design the core architecture with this in mind. Figure 3 describes this idea.

An optimized CPU architecture performs the 90% load, while a distributed processor filters out the impulse functions. The distributed processor can be general-purpose media access control (MAC), a DSP, or an MPEG II/IV accelerator. It's targeted to filter out the IPR high-overhead requirements for a partic-

ular, unique market segment. Designs optimized in this way eliminate the baggage associated with high-MHz, general-purpose PC designs.

Historically, product cycles follow an interesting trend. Trips to see a play or theatre evolved into trips to view projection movie screens, to televisions at home, then to VCRs. Now, delayed or replay TV allows complete personalization. Trips to the orchestra evolved to phonographs at home, to radio, then to personal tape/radio players. Currently, MP3 players provide flexibility and portability.

Computers began as room-size mainframes limited to the select techno elite for access. These evolved into minicomputers, workstations, desktop PCs, and laptop PCs. Now, the PDAs and dedicated IA devices provide data portability and targeted functionality. Products tend to progress from large, general-purpose, impersonal static forms to portable, personal, flexible, market-targeted forms.

Conflicting requirements

As Internet ubiquity and personalization drive the requirements of future CPU and system design, a few fundamental cornerstones for technology and architecture will emerge:

- battery life,
- portability,
- security,
- connectivity,
- user interface,
- application compatibility,
- universal data access, and
- cost.

This list of requirements presents an enigma for the CPU and system architect. While battery life, portability, and cost require simple, application-specific solutions, previously described performance impulse requirements, universal data, security, and user interfaces require the ability for higher performance. Of these requirements, the user interfaces present the most conflicting requirements for system design specifications.

The consumer and the usage model dictate a more personal interaction with the CPU and system. Today's keyboard and mouse evolve into voice command and control, voice recognition, handwriting recognition, fingerprint identification, and other biometrics. These requirements call not only for specific digital performance requirements but also for specialized analog capability to permit better interaction with the analog-centric human user.

CPU and system optimization for one set of requirements will cause unacceptable design trade-offs in other areas. For example, architecture cannot be designed solely for high-overhead general-purpose performance, or it will sacrifice battery life, portability, and cost. The process, voltage, and design issues described earlier along with the requirements of the Internet and consumer usage model will bring about a divergence point that will require changes in CPU and system design philosophy.

Figure 4 describes the diverging requirements in the consumer marketplace for performance, user experience, and connectivity. Sufficient performance provides a better experience to the novice user. However, the highest performance system will provide no more perceived performance to that inexperienced user. As indi-

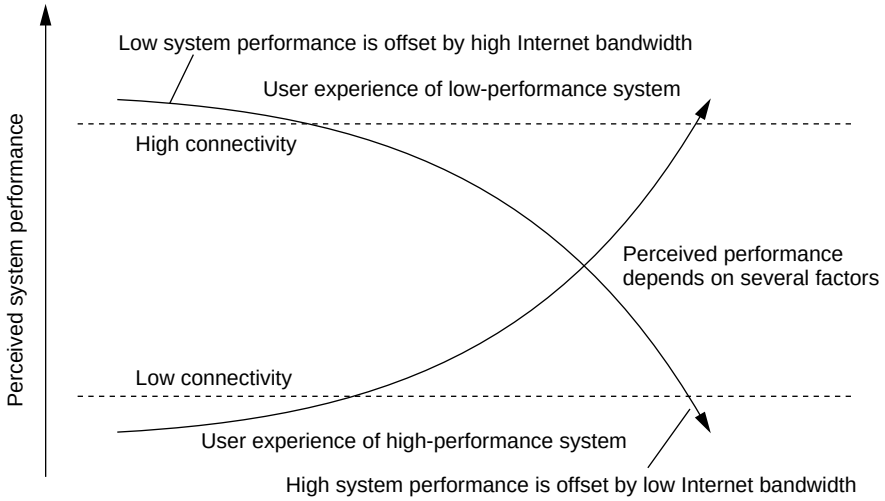

Figure 4. Diverging requirements. User experience depends on a balance of system performance and connectivity bandwidth.

cated earlier, connectivity is a key element in the system architecture. A user with the highest performance system and a poor connection will perceive the usage model of a person with a very low performance system and high-bandwidth connection.

These issues are occurring because the required pace of technological innovation is increasing, while the actual time to market and effectual usage of technology is decreasing. Human nature being what it is, we react skeptically and negatively to change. Technology is no exception. The current installed base of more-transistors-and-faster-MHz designers is comfortable with the status quo. The infrastructure to support the current design path is large and heavily capitalized. Software companies, original device manufacturers (ODMs), OEMs, and users are all familiar with the current architecture trend. They may acknowledge the need for change but lack the energy required to overcome "rear view mirror" inertia.

New architectures

Paradigm shifts, however, do occur—either with subtlety or with ferocity. The mainframe or supercomputer business is an example of necessary change manifesting itself in new system architectures. The old school architecture of supercomputers used a proprietary-core CPU and systems with the sole goal of maximizing MHz or computation cycles. Proprietary operating systems, applications software, design tools, and the lack of a large knowledge base were secondary to the performance-at-all-costs objective.

Designers of today's systems use off-the-shelf x86 or PowerPC processors. Brute force has given way to flexibility and time-to-market needs. The original architectures couldn't continue to meet the market demands. Costs and technology-to-market times increased past the acceptance point. Necessity dictated the paradigm shift to distributed processing over multiple smaller, easily producible CPUs.

Performance flexibility and application flexibility are a result of the new architecture. Each machine can be tailored to a certain performance level by removing or adding processing nodes. Infrastructure leverage is another result. These architectures now embrace the knowledge base of PC and workstation CPU architects, software writers, and manufacturers.

Products tend to progress from large, general-purpose, impersonal static forms to portable, personal, flexible, market-targeted forms.

The PC's success has largely been a result of a standard, open platform. As the new Web-based economy grows, different platform requirements will evolve or be created. Personalization implies differentiation of platforms and technologies supporting those platforms. As biometric, genetic, and chemical technologies evolve, they will begin to be incorporated with CPU designs. Personalization, flexibility, and quick time to market will dictate a quick-turn design methodology.

This is in direct opposition to the design style and methods incorporated in today's CPUs. Time to market for new architectures is measured in three-year increments, while spins of new processors based on known designs occur every six months. This design method requires that the architecture be correct for up to six years in the market—three years for development and three years in production. The market requirements will no longer accept the current long pipelined process design cycle. The Internet has broken down all communication and knowledge barriers. This breakdown has increased the worldwide productivity, manufacturing, knowledge, and design base. Design cycles must decrease or become the bottleneck for future progress. A more robust and flexible architecture is required.

New consumer model

Consider the system architecture philosophies in other market segments. The automotive sector lets consumers either buy standard models available on a dealer's lot or order a personalized feature set from an available menu. Customers can drive off of the lot the same day or wait six to eight weeks for the

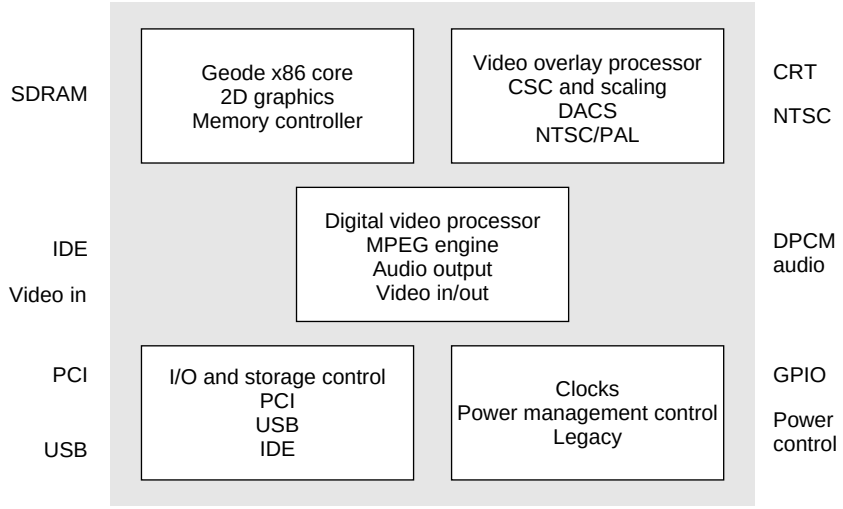

Figure 5. Future multimedia system on chip.

build-to-order version.

The housing market lets customers buy a predefined feature set in an existing home or wait three to six months to specify almost all details in the home's construction. Given the desire for personalization of products and the time value given by consumers for personalized products, architectures in the Internet-driven millennium must accommodate build-to-order personalization of computers, microcontrollers, or systems.

To realistically achieve this model, designs must permit consumers to access a menu of features and functions available to optimize and personalize the final product. Those desiring a lightweight hand-held appliance will optimize for battery life and perhaps wireless access. Others building a home Web browser might choose to select ADSL and an x86 CPU core for Internet compatibility. Still others might choose to add a distributed processor for MPEGII/IV acceleration to an x86 core for an optimal browser with multimedia capability (see Figure 5).

The a la carte model for product personalization will obviously place many burdens on the CPU and system architect. But, these aren't insurmountable. Architects will use SOC techniques to provide quick builds and time to market for customizable solutions. A building-block approach using presimulated, pretimed, prebuilt modules will allow for cycle times gated only by the back end of the manufacturing process.

Another requirement to achieve this mode of product distribution is for system and CPU architects to take advantage of the software infrastructure and flexibility allowed by open-source offerings. SOC and distributed-processing chip-building techniques are historically not the cause of production delays. Software availability, support, and knowledge base are the bane of product schedules. However, necessity will demand that these and other issues are solved, and consumers in the new millennium will purchase build-to-order products optimizing features and functions from an a la carte menu of items.

Performance and the Net

The dichotomy of requirements in the CPU and system design world is the need for a base level of performance and the need to support impulse functions representing high levels of performance requirements. With the exponential growth of the Internet and its spread to other media such as cell phones, PDAs, and appliances, computation power is growing at almost factorial growth rates. The puzzle is how to make use of this power. Efforts are under way today to use this interpute (Internet compute) power.

For example, the SETI (Search for Extra Terrestrial Intelligence, http://setiathome.ssl.berkeley.edu/) project allows users to make their processors available as part of a group effort to analyze satellite and radio telescope data in the search for extraterrestrial intelligence. Likewise, future microsatellite system designs will provide small, lightweight, low-processing nodes that work in consort to solve problems requiring higher processing and functional capability than any individual microsatellite can provide.

At a local level, many programs and applications already make use of available computer power to help with high-performance applications. Workstations and enterprise servers using mesh networks of x86 processors are examples.

Computer users in the 21st millennium won't need to carry or have available high-performance computation power. Users will access the available interpute power to handle performance-driven requirements. This use of interpute power instead of local CPU MHz will cause the CPU and system architect to optimize for data packet transfer and local user interface capability. These requirements match those of the Internet appliance device architecture.

Personalized Internet appliances will be data and user interface conduits to the end user. The desktop PCs, workstations, and enterprise server farms sitting on the Web will provide the interpute power needed for high-performance requirements. From a system architecture point of view, the key elements required to achieve this meshlike architecture are data standardization and application transcoding. Data must be recognizable by all nodes on the mesh, and applications must be transcodable or changeable from one to another for the compute power available on the Internet to be usable as interpute power to the end user. Standards bodies like the IEEE and the WWW consortium must work to achieve these standards.

Human interaction

Finally, CPU and system architectures will evolve to achieve more and better interaction with each other and with humans. At the local level, analog technological improvements will allow for CPUs to communicate directly with the nervous system to aid in medical, physical, and mental improvements. At the Internet mesh level, systems will take on the capabilities of service or informational agents. Interpute power will allow for humanlike features such as emotion, anticipation of events or results, adaptation to events, and ability to communicate. Users will interact with personal agents to aid with problems or provide advice for medical, travel, emotional, investments, and many other areas.

CPU, microcontroller, and system architectures in the 21st millennium will become more flexible for market and customer requirements. The Moore's law progression of technology flies against the historical model of personalization and market-specific requirements of consumer products. Distributed processing and system-on-chip design techniques will allow for a la carte personalization of products. Analog advancements will enhance the user interface experience of personal Internet appliance devices and allow for interaction at the neural-network level. The standardization of data, transcoding of applications, and explosion of computation power on the Internet will let users take advantage of interpute power for high-performance applications. Necessity dictates change. Divergence points dictate change. Both have brought architecture to the cusp of required invention. MICRO

Design cycles must decrease or become the bottleneck for future progress. A more robust and flexible architecture is required.

Chris Herring is director of strategy and architecture for National Semiconductor's Information Appliance Group. He is responsible for business segment analysis, strategic initiatives and alliances, and product architecture and road maps. Earlier, he served as National's director of platform development, managing the design of general-purpose and segment-specific platforms such as the National Geode WebPAD, thin-client, and set-top box reference designs. He also worked with semiconductor and systems design, architecture, and management working for IBM, Cyrix Corporation, and National Semiconductor. He has designed numerous memory and cache controllers; interface buses; and cores for mainframe, AS/400, and PC architectures, holding 17 patents in these areas. Herring received BEE and MSEE degrees from Georgia Institute of Technology.

Direct comments about this article to Chris Herring, National Semiconductor Corp., IA Group, 2302 Lake Park Drive, Longmont, CA 80503; chris.herring@nsc.com.