■ VIEWPOINT

Integration Challenges PC Architects

As Transistor Counts Head for 100 Million, New Designs Are Needed

by Mike Webb, President, Summit Research

Silicon advances of the 1990s have put increasing pressure on the PC architecture standard created in 1981. The old PC peripheral chips were the first to go—what design today uses the 8237 DMA controller still present in PC chip sets? VGA video gave way to SVGA, then to SVGA with increasingly sophisticated Windows and multimedia support. The ISA bus remains the least common denominator for peripheral connection, but new peripheral cards seeking more bandwidth use PCI.

With IC fabrication nearing the 0.25-micron mark, chips will contain tens of millions of transistors, but the basic PC architecture uses about 4.2 million transistors (not counting memory chips) for the entire system. PC architects' ability to use these transistor budgets effectively may determine the PC's long-term viability as a computing standard.

The PC's rapid adoption has been spurred by rising computing performance, lower computing cost, and the ability to process new data types (e.g., multimedia) to improve human interaction. Increasing integration supports all three of these areas. For the microprocessor, higher transistor counts allow more execution units to operate in parallel and larger on-chip caches to feed them.

Intel and others have also concentrated on removing performance bottlenecks near the processor caused by legacy hardware from the original PC architecture. The result has been initiatives such as PCI, MMX, AGP, and the Pentium Pro bus architecture.

The rapid adoption of multimedia PCs for home and business has changed the peripheral landscape, providing niches for smaller players. Peripheral-chip vendors in the audio, video, communications, and disk-control arenas have stepped in to fill the gaps in interface performance left by Intel. PC peripherals that offload main CPU graphics tasks are now commonplace. Audio processors and modem DSP chips are taken for granted.

Superintegration poses a problem. Integrating a CPU, first- and second-level caches, and separate processors for audio, video, and communications onto one chip may soon be possible, but such a product may not be viable. Instead, it may make more sense to use these millions of transistors to add features to current designs or to reduce the overall cost of the system.

Transistor Budgets Spiral Upward

Integration to improve microprocessor performance is still driving single-chip transistor budgets upward. Advanced microprocessors and DSPs provide the funding needed to develop and deploy the most advanced processes. The Pentium Pro CPU, for example, contains 5.5 million transistors in a 0.35-micron technology; its companion 512K SRAM consumes 31 million. TI claims its 0.25-micron CMOS process, due in production next year, will allow transistor counts as high as 125 million *(see* **1008MSB.PDF** *)*. This claim may be somewhat aggressive, but most projections show 100-million-transistor logic/memory devices available by the end of the decade.

Thus, there will be plenty of transistors for any task you want to move onto the CPU chip. But history indicates that collapsing existing PC systems onto a single chip may not be the best approach. As it becomes feasible to integrate a "standard" PC in one or two chips, the standard PC definition moves up a notch as people demand more from their personal computers. Remember the AMD 286ZX and the Intel 386SL? Both condensed a standard PC into one or two chips plus memory. Both were introduced less than five years ago, but neither had much success in the PC market. So superintegration isn't likely to collapse all PCs onto a single standardized chip.

In the fast-moving silicon industry, companies are already pursuing many other paths to capitalize on integration. Cost reduction alone has never been enough of a reason to integrate PC components. Devices that lower system cost while adding user-visible features have done much better in the market. Cirrus and S3 have innovated with Windows graphics accelerators followed by multimedia video accelerators. In contrast, the company with the lowest-cost VGA controller just isn't around anymore.

Entry-level systems often lean more toward cost efficiency, with as many advanced features as the budget allows. Feature differentiation means more in high-end PC systems. With these lessons in mind, let's take a peek at how future transistor budgets may be used to advance the performance of tomorrow's PC.

GigaChips Combine Processor, Peripherals

Terms like processor or peripheral may be inadequate to describe these upcoming devices. Let's call them GigaChips. GigaChips will likely have processors, memory, and peripheral logic integrated in virtually every device. Localized resources avoid the delays involved in going off chip. It seems Intel has foreseen this possibility in the P6 bus architecture by supporting up to four processors. This multiprocessor support is targeted mainly at servers today, but as devices like S3's Trio64UV+ continue to add outbound processing for

Feature	Transistors
Main CPU Eight-issue nonblocking parallel execution units (using P6 core as basis)	25 million (8× P6 core plus interconnect)
L1 cache 512K with 256K I-cache & 256K D-cache	50 million
L2 cache controller with tags	5 million
Peripheral bus interfaces New high-speed interconnect	1 million
Main-memory interface	1 million
Compatibility logic Just a drop to pull legacy apps	250K
Video processor & transform engine	5 million
Advanced DSP engine Audio/video/speech/data	5 million
Control logic, cache coord, internal bus	5 million
Total transistor count, GigaChip One	**97.25 million**

Table 1. GigaChip One combines a high-performance processor with cache and system logic.

audio and video data streams, these devices may migrate toward the main CPU bus for enhanced performance and cache coordination.

Let's start building our hypothetical GigaChip system by assuming software integration issues are magically solved by someone and the hardware building blocks are available for whatever functions we would like to integrate. Later we will look at obstacles that might arise. With 100 million transistors to spend, let's go shopping!

The GigaChip shown in Table 1 approaches the integration question from the CPU's viewpoint. Pulling many high-bandwidth system functions on chip makes sense, because it increases performance by eliminating the delay required to go off chip. As clock speeds rise in the GigaChip era, penalties for off-chip accesses increase. Bus speeds will not increase proportionally, as setup times, arbitration, and hold times are beginning to dominate interchip communication. It's easy to imagine a major shift away from low-level bus protocols once on-chip caches grow large enough to disguise the overhead of packet-based communications. Communication between GigaChips stands to gain from packet-oriented small-signal-swing buses. This change allows longer, faster bursts of data between devices once communication is established.

Feature	Transistors
Main CPU 64-execution-unit array RISC core with scheduler	32 million
Translation pipe for x86 to RISC Optimized for parallelism	20 million
DSP array Eight general-purpose DSPs	16 million
CPU array L1 cache 512K with 256K I cache & 256K D cache	50 million
L2 cache controller with tags	5 million
Peripheral bus interface New high-speed interconnect	1 million
Main-memory interface	1 million
Compatibility logic Just a drop to pull legacy apps	250K
Control logic, cache coord, internal bus	5 million
Total transistor count, GigaChip Two	**130.25 million**

Table 2. GigaChip Two is designed for maximum performance with a 100-million-transistor budget.

If something like GigaChip One becomes reality, other component vendors, as well as system vendors, will need close engineering relationships with the GigaChip designers to capitalize on the sophisticated interfaces needed between this device and the rest of the system. Intel's MMX extensions may pave the way for multimedia processing to move onto the main CPU. GigaChip One provides complete audio, video, and communications functions concurrently with the data-processing functions of the main superscalar engine.

The thermal performance of upcoming deep submicron processes is critical to realizing such a chip. On-chip thermal problems require significant voltage reductions. The 200-MHz Pentium Pro dissipates 28 W (typ) at 3.3 V. Even with a supply-voltage reduction to around 1.8 V, the GigaChip is likely to dissipate 50 W or more.

Packaging will continue to grow in importance in allowing the silicon to reach its peak performance. Rapid heat dissipation is necessary for GigaChips. Packaging must also provide low-impedance, low-inductance connections to the rest of the system as clock rates continue to increase. These improvements are needed to minimize the performance loss in off-chip connections.

Pin counts for our hypothetical GigaChip are worth examining. The system partitioning reduces the need for high-pin-count buses, but wide main-memory and peripheral buses are needed, along with their related arbitration and control logic. Several hundred pins seem likely, well within the range of today's BGA packages.

Packet-oriented buses could reduce pin count. The intelligence required at the other end might allow the main CPU to hide most of the latency getting on and off the off-chip buses. Another way to reduce the pin count would be with narrow, fast interfaces like Rambus.

Maximizing CPU Performance

Perhaps GigaChip integration will lead in a different direction. GigaChip Two, shown in Table 2, assumes the CPU remains the CPU without trying to integrate all the other peripherals. As usual, the CPU designers ran somewhat over the transistor budget. This approach aims to maximize performance for both the DSP and data-processing functions while minimizing the specialization of the blocks as much as possible. If this approach works out, it has several benefits.

GigaChip Two is significantly easier to design and test, due to its less specialized cores. Development time is already a limiting factor on new CPU designs. This device has the

added benefit of being much more scalable. When 2010 rolls around and we are wondering what to do with a billion-transistor budget, we will be well prepared.

Both these GigaChips use most of the newly available transistors for memory, to get as much data and code as possible close to the execution units. This method for increasing performance is well proven and should continue to be used in high-performance segments of the PC market.

Integration to Reduce Cost

GigaChips for other PC market segments may use different approaches. The most cost-sensitive PCs are likely to use this technology to optimize for cost rather than performance. A good rule of thumb has been that devices of about 50% of the maximum die size tend to be very cost effective. Table 3 shows a device using about two-thirds of the transistor budget of GigaChip One to build an entry-level machine.

This device surrenders L2 cache support for die size. I suspect our hypothetical third chip would really have most of the performance of GigaChip One at a much lower cost. Single-chip transistor budgets in the next five years may actually surpass the needs of any reasonable PC configuration that simply extends today's features. The challenge for chip and system designers may be in finding new functions that use these transistors effectively.

For example, 3D graphics combined with real-time audio and video communications may be just the ticket for the consumer PC. Video integration might include the frame buffer as well, but the drawback here will be that the most efficient frame buffers use DRAM processes, while CPUs are predominantly SRAM-oriented processes.

What if the PC becomes more of a communications center both at home and work? The transistors will be there for this approach as well. The communications PC might focus more transistors on on-the-fly compression/expansion of simultaneous voice/data/video to provide highly interactive communications, close to what you would expect between two people in the same room. GigaChips may lead from the Personal Computer to the InterPersonal Computer.

Organizational and Business Obstacles

Whether GigaChips proceed according to this speculation or in yet another direction, one thing is sure: the overhead of the 1981 legacy logic will continue to diminish to insignificance. This burden, which has constrained the x86 architecture from competing in some segments of the market, will shortly become irrelevant.

Simultaneous innovations in the PC architecture along the way have laid a reasonable foundation to support 100-million-transistor PC chips. The P6 multiprocessor architecture, advanced graphics, and digital signal processing for audio, video, and communications provide base-level high-bandwidth architecture enhancements that are ripe for further integration.

Feature	Transistors
Main CPU Eight-issue nonblocking parallel execution units (using P6 core as basis)	25 million (8× P6 core plus interconnect)
L1 Cache 512K with 256K I-cache & 256K D-cache	25 million
Main-memory interface With intelligent working-set exchanges	5 million
Compatibility logic Just a drop to pull legacy apps	250K
Video-stream processor engine	5 million
Compression/expansion engine Audio/video/speech/data	5 million
Control logic, cache coord, internal bus	2 million
Total transistor count, GigaChip Three	**67.25 million**

Table 3. GigaChip Three is designed for a cost-effective die size and moderate performance.

Organizational and business obstacles may become the gating factors. The rapidly changing dynamics of PC systems provide challenges in the product definition of GigaChips. Except for the largest chip vendors, companies will need to cooperate in designing these complex devices, providing the flexibility necessary to allow system differentiation while pushing the standard PC configuration forward.

GigaChip One, for example, is a logical extrapolation from today's Pentium Pro, TMS320-family DSP, S3 video accelerator, and C-Cube audio/video encoder/decoder. The system technology, chip architectures, and software expertise required to deliver such GigaChips may push semiconductor suppliers to new levels of cooperation. On the other hand, perhaps the task will prove too daunting for any but Intel, TI, and a few others.

Perhaps the users of such devices will demand new business and marketing approaches from most silicon suppliers. Just as the old second-source arrangements demanded by system companies in the early 1980s have disappeared, this new era of integration may give rise to second sourcing of silicon building blocks between semiconductor houses, so they can keep up with the market for superintegrated devices. The winners may become the companies with the best portfolio of building blocks to trade and the most sophisticated design environment in which to master the integration and testing of the resulting GigaChips.

GigaChips, like most steps along the silicon integration trail, offer the promise of incredible new systems. Beyond the technical challenges of creating such devices lie new business and organizational challenges. Innovation in business partnerships and organizational teamwork will unlock the promise of GigaChips in the next decade. M

Mike Webb is president of Summit Research (Austin, Texas), a consulting firm specializing in strategic market research and business planning for the semiconductor and computer industries. He can be reached for questions or comments at Si2Au@aol.com.