Cahners
MICROPROCESSOR REPORT
www.MPRonline.com
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

Processors Penetrate Gigahertz Territory

Compaq, IBM, and Intel Describe 1GHz Processors at ISSCC

By Keith Diefendorff {2/28/00-02}

At this year's International Solid-State Circuits Conference in San Francisco, where the industry's top semiconductor gurus assemble each year, the hot topic on the microprocessor front was frequency. Compaq, IBM, and Intel each presented papers describing operational 1GHz processors. Not presenting a paper (but not wanting to let Intel steal all the gigahertz glory) AMD showed up to demonstrate a 1.1GHz Athlon. Although it didn't quite reach the 1GHz threshold, Motorola presented a paper on a 780MHz G4—not bad for a short-pipeline processor.

Compaq, Samsung Push 21264 to 1GHz

To boost the Alpha 21264 (see *MPR 10/28/96-02*, "Digital 21264 Sets New Standard") from its current speed of 700MHz to 1GHz, Compaq and Samsung shrank its transistors to 0.18 micron, leaving the metal stack at 0.25 micron but adding a seventh layer of metal. This approach left the die size unchanged (193mm^2), allowing the current 587-pin PGA package to be reused. At 1GHz, the new 1.65V part dissipates 65W and will trounce all other known processors if it meets Compaq's estimates of 60 SPECint95, 110 SPECfp95 performance. (From these figures, we estimate baseline SPEC scores of 54 and 91 respectively.)

To maintain a reasonable leakage current (I_{doff}) in the faster part, Compaq and Samsung were unable to scale the threshold voltage (V_t) of the transistors by the same amount as the supply voltage (V_{dd}). Thus, to recover the speed lost to a relatively higher V_t, the companies introduced a low-V_t device, 80–100mV below nominal, which boost drive current (I_{dsat}) by 18%. These devices were used judiciously in critical timing paths only, but were not used in dynamic circuits because of their susceptibility to noise and their higher leakage currents.

While the new 21264 runs at 1GHz in its wire-bond PGA package, Compaq and Samsung hope to boost the processor's speed another 8–10% (80–100MHz) by flip-chip mounting the part. The 65W part draws 40A, which the companies say creates a 320mV worst-case voltage drop between the pins and the center of the die. With flip-chip mounting, this number drops to 125mV, which should allow the part to run at even higher clock rates.

AS/400 Speeds Up 50%

IBM's AS/400 group from Rochester (Minn.) showed the fruits of IBM's silicon-on-insulator work (see *MPR 8/24/98-02*, "SOI to Rescue Moore's Law"), using it to bump its Pulsar processor from its current 450MHz in 0.22-micron bulk CMOS-7S to 550MHz in the same process on SOI and again to 660MHz in 0.18-micron CMOS-8S2SOI.

The new 8S2 version of Pulsar, code-named Sstar, also includes an on-die L2 directory with nearly twice the capacity of the original's, taking the transistor count from 34 million to 44 million. Despite the higher transistor count and higher frequency, the 1.5V 8S2 part has an 8% smaller die (128mm^2) and, at 18W, dissipates almost 20% less power.

IBM has apparently learned to control the tricky floating body of SOI transistors. Since the body of an SOI transistor is electrically isolated, the body-to-source voltage can vary, altering the V_t of the devices. Not only has IBM learned to deal with the floating body, it has found it can leverage the effect by contacting the body and driving it to different levels. Using this technique, IBM says it can speed up a controlled-buffer chain by as much as 18%. There is an area tradeoff for contacting the body, but when used in critical paths, the technique can improve performance even more than the natural speedup from SOI.

Mainframes Want Frequency Too

Struggling to keep pace with PC processors, mainframe processors are joining the frequency race. IBM's Poughkeepsie-based S390 group described a 760MHz G6 processor, which is about 20% faster than the 637MHz processor it shipped in a 12-way SMP configuration in June of last year. The new G6 processor uses IBM's 0.22-micron CMOS-7S process, making it the world's first commercial processor to use copper. The 25 million transistors on the 215mm^2 G6 die burn nearly 40W at 1.9V.

The G6 is a single-issue design with a short, seven-stage pipeline and 256K of on-die L1 cache. A fast low-V_t device is used in critical speed paths. IBM said that 10% of the transistors are low-V_t devices, and 97% of the devices in the critical paths are low-V_t devices. The company said the use of these devices boosted frequency by about 10%. Apparently the company debated using dynamic circuits rather than low-V_t transistors. Although dynamic circuits would have been slightly faster, they would also have used considerably more power. Since heat buildup was the primary factor limiting the frequency of the processor, IBM elected to go with the low-V_t devices, which added only 0.5W to the part's power consumption.

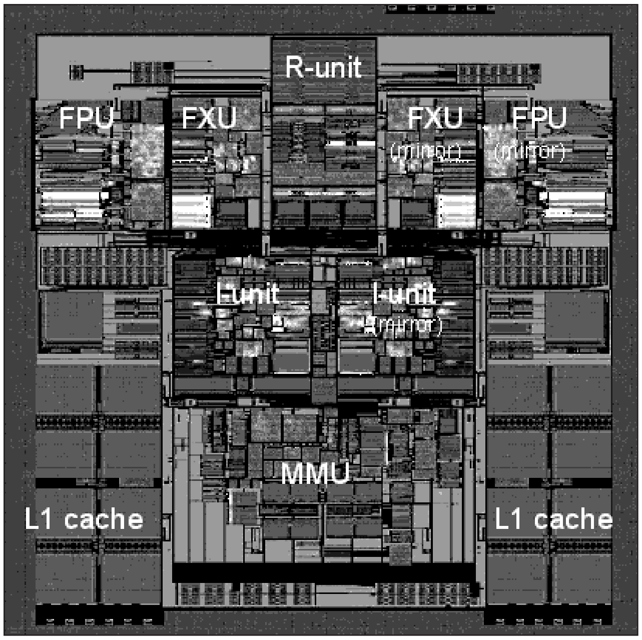

Figure 1. IBM's G6 mainframe processor uses duplicate execution units and error-protected MMU to detect errors. The G6 is built in IBM's 0.22-micron copper CMOS-7S process and contains 25 million transistors in 215mm^2 of silicon. (Source: IBM)

Copper interconnects also contributed to the speed of IBM's processors by reducing RC delays by 30%. Simulations show that copper wires allowed IBM to reduce the wire-delay component of critical-path cycle time from over 9% in the CMOS-6X G5 processor to below 8% on the CMOS-7S G6. This fact is more significant than it may seem because, had IBM stayed with aluminum wire, the RC-delay component would probably have increased by several percentage points with the denser packing and higher transistor speed of the CMOS-7S process.

An impressive feature of the G6 is its redundancy, which is apparent in Figure 1. The processor uses duplicate instruction and execution units to detect errors; other units (such as the MMU and L1 cache) are fully protected by ECC and parity. If any discrepancy is detected during run time, the transaction is retried in an attempt to get past soft errors. On a second failure, the error is assumed to be hard, and the state of the processor is dynamically moved to a good processor.

Even more impressive is the G6 system's packaging. A 14-way SMP system, including 31 chips total, is packaged in a single MCM with 2,662 signal I/Os. The MCM houses 1.4 billion transistors on a 5-inch-square substrate, and it burns a kilowatt of power. The MCM is cooled with a 0°C chiller that lowers junction temperatures to 10°C.

IBM Runs Short Pipe at 1GHz

To reach 1GHz, designers are resorting to longer and longer pipelines. Intel's Pentium III has a 12-stage pipeline, for example, and even longer pipelines are being considered for future processors. Unfortunately, long pipelines dramatically increase the complexity of processors, as considerable additional logic, such as branch predictors and fast operand-forwarding paths, must be added to retain some modicum of pipeline efficiency. Bucking the long-pipeline trend, IBM decided to build an experimental PowerPC processor in its Austin Research Lab to determine if it was possible to run a short pipeline at high frequency, thus avoiding the extra complexity of long-pipeline designs.

Apparently, it is possible. IBM succeeded in building a six-stage pipeline processor that operates at 1.0GHz in IBM's standard 0.22-micron six-layer-copper CMOS-7S process. At the fast end of the process distribution, the part actually clocks in at 1.15GHz, burning 112W at 1.87V (101°C). IBM says that further frequency increases are limited by heat dissipation, but it also said power on the chip is double what it could have been had the part been designed by more than just a few people in a research lab. The small design staff required some compromises that were time efficient but power inefficient: For example, only 64-bit- and 4-bit-latch macros were used, resulting in many nonfunctional power-hungry latches in the design.

The speedy design is a full 64-bit PowerPC processor, with 19 million transistors occupying 150mm^2. It's a single-issue design with a single-cycle integer ALU, a two-cycle pipelined load unit, and a four-cycle fully pipelined FPU that supports fused multiply-add. Branch-miss penalty is five cycles, and a simple history buffer handles exceptions late in the pipeline. Delayed-reset and self-resetting dynamic circuits are used throughout the design.

An unusual feature of the part is its enormous on-chip decoupling capacitor. As Figure 2 shows, the capacitor occupies approximately 35% of the die. IBM said, however, that the capacitor is not really necessary, as there is no noticeable difference in operation with the capacitor disconnected. Since it was a research vehicle, extra die area was available. The designers filled the extra area with the capacitor, just in case power-integrity problems arose.

IBM did not disclose the performance of the machine, but, as a research vehicle, the part is not likely ever to see the light of day in an IBM server or an Apple Macintosh. The experiment is interesting because it proves that high frequency can be achieved with short pipelines. IBM didn't say whether the techniques it developed for this design would scale to even higher frequencies with longer pipeline designs or could be used for superscalar designs. We suspect, however, that many of the tools and circuit-design tricks used in this design are being applied to IBM's next-generation POWER4 processor (see *MPR 10/6/99-02*, "POWER4 Focuses on Memory Bandwidth"), which will be deployed in commercial servers in 2001.

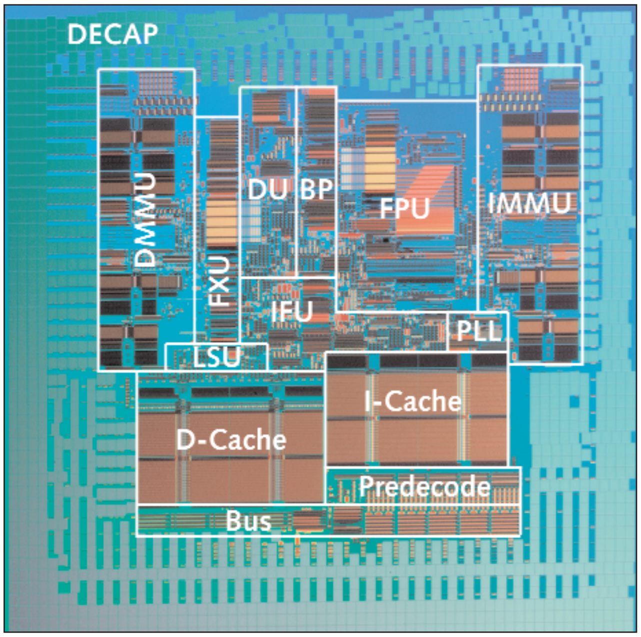

Figure 2. This PowerPC processor from IBM Research bucks the trend toward long pipelines, but it still runs at over 1GHz. The 19-million-transistor processor occupies 150mm^2 of silicon, almost half of which is decoupling capacitor. IBM says the capacitor isn't really necessary. (Source: IBM)

Noncopper Coppermine Hits Gigahertz Mark

Intel devoted most of its ISSCC presentation to rehashing the features of its Coppermine design (see *MPR 10/25/99-01*, "Coppermine Outruns Athlon") and to defending the company's decision to forgo copper interconnects in its 0.18-micron process. It was not clear from the presentation what, if anything, was done to the currently shipping 800MHz Coppermine, shown in Figure 3, to boost it to 1GHz. Nearly any shipping 800MHz processor could be deep-sorted to identify a few that run at 1GHz at room temperature and nominal voltage (assuming, of course, that the part's cycle time is not solely determined by wire delay, which is rarely the case).

At least part of the speedup is probably due to the use of the notched-poly process that the company described at the International Electron Devices Meeting (IEDM) last December. Otherwise the process is apparently Intel's standard 0.18-micron P858 process (see *MPR 1/25/99-06*, "Intel Raises the Ante With P858"). Notching the gate poly is an exceptionally clever idea: by undercutting the gate, Intel can heavily dope the source-drain regions (for low resistance) without creating a high gate-overlap capacitance. The result is transistors with remarkably high I_{dsat} and therefore speed. Intel showed a shmoo plot indicating that at the top of the P858 voltage range, 1.7V, the Coppermine part will actually operate at 1.1GHz (at room temperature). At 1.45V and 1.0GHz, Intel said, the processor dissipates over 30W.

Once worried about the ability of its Socket 370 PGA package to handle high frequency, Intel now says it is getting better performance out of that package than from the more-expensive SECC2 (single-edge-contact cartridge) module.

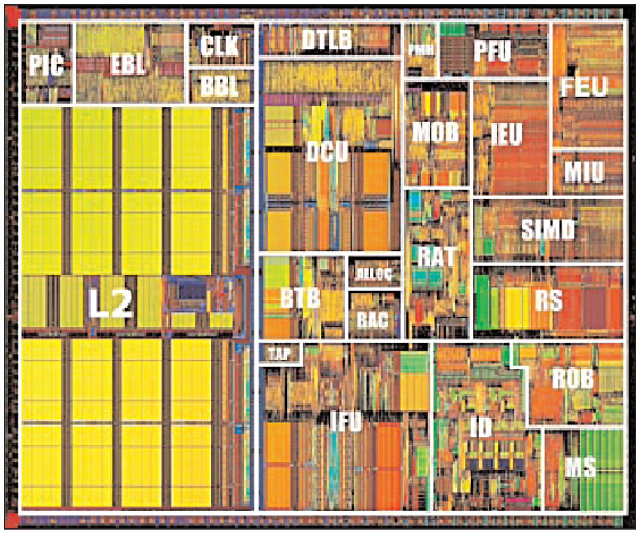

Figure 3. Intel's 018-micron Coppermine Pentium III processor is currently shipping at 800MHz. The chip implements 23 million transistors in 106mm^2 of silicon. The processor presented at ISSCC is apparently similar but uses Intel's notched-poly transistors to achieve a speed of 1GHz. (Source: Intel)

The Socket 370 package uses an organic fiberglass substrate with decoupling capacitors mounted directly beneath the chip on the opposite (pin) side of the substrate.

These days, Intel seems to be spending an inordinate amount of airtime defending its decision to delay copper interconnects until its 0.13-micron generation. That decision runs contrary to the decisions of most other semiconductor manufacturers, which are charging ahead, as fast as their little budgets can carry them, to integrate copper into their 0.18- and 0.15-micron processes. At the ISSCC presentation, Intel again took the opportunity to defended itself, arguing that by applying incremental design effort to only 3% of the chip's circuits, and by using repeaters in some of the long wiring runs, it was able to create a design whose cycle time is dominated by transistor delay and is therefore nearly independent of interconnect technology. Using this technique, Intel says it achieved the same speedup as it could have with copper wires, while eliminating the risk and cost of introducing copper into its fabs at this time.

This argument, however, fails to acknowledge that if Intel had used copper in addition to the incremental work on 3% of Coppermine's circuits, even higher speeds might have been achieved. It is also not clear that designing critical paths to be transistor dominated necessarily yields the fastest possible circuits.

AMD, Intel Play Leapfrog

Determined not to be outdone by Intel, AMD appeared at ISSCC to give a backroom demo of its newest Athlon chip, code-named Thunderbird, running at 1.1GHz. The new chip uses the existing Athlon core but adds an on-chip L2 cache. AMD did not disclose the cache size, but we expect it to be 256K. Thunderbird is implemented in AMD's new copper 0.18-micron HIP6L process that it originally licensed from Motorola (see *MPR 8/3/98-en*, "Motorola, AMD Swap Technology"), and the part was actually built in AMD's recently opened Fab 30 in Dresden.

We expect this game of leapfrog and one-upsmanship between Intel and AMD to continue into the foreseeable future. In fact, no sooner had ISSCC ended then Intel jumped ahead once again with a demonstration at its own Developer Forum (IDF) of a 1.5GHz Willamette. This processor is based on Intel's next-generation microarchitecture, which has a pipeline twice as long as that of the Pentium III as well as advanced features such as a trace cache and a unique double-pumped ALU (see *MPR 2/28/00-03*, "Quicktake: Willamette Revealed"). The Willamette that Intel demonstrated at IDF was built in the same 0.18-micron process as the 1.0GHz Coppermine, so, while many technologists may agree that Intel's choice of aluminum interconnect was not the best, it is obviously not a mistake so large that it is preventing the company from building very high-frequency parts. And, as long as Intel can stay in the frequency lead, the issue of copper or aluminum seems moot.

Intel did not say exactly when Willamette would enter volume production or at what speed. We expect, however, that production will begin in 4Q00 at 1.1GHz, reaching 1.5GHz production next year. Intel did say, however, that Coppermine would ship at 900MHz before midyear and at 1GHz sometime in 2H00.

Moto, HP Miss Gigahertz Target

Motorola presented a paper on a 780MHz 0.18-micron six-layer-copper HyperMOS 6 (HIP6) PowerPC processor, the same processor it first described at last year's Microprocessor Forum (see *MPR 10/25/99-02*, "PowerPC G4 Gains Velocity"). The new G4+ part is a variation of the G4 processor currently shipping in Apple Macintosh G4 systems.

The new design adds two stages to the G4 pipeline, increases the instruction-issue width from two (plus a branch) to three (plus a branch), and ups the number of execution units by two. Other enhancements were made to retain the same pipeline efficiency and IPC (instructions per cycle) as the initial five-stage G4 design. The G4+ also adds an on-chip 256K L2 cache while retaining the external-cache interface (and on-chip tag directory) as an L3. The additions bring the transistor count of the new design to 33.1 million: 6.8 million transistors in logic, 26.3 million in SRAM. The die has 286 C4 solder balls for signals and 755 for power and ground. At 105mm^2, the die is comfortably small from a manufacturing-cost perspective.

Motorola did not say when it would ship the new G4+, but Apple probably hopes it is soon. With the current G4 mired down for months at 450MHz, Motorola really needs to boost PowerPC's frequency. After prematurely announcing a 500MHz G4 in September of last year, Apple has spent the intervening months waiting for Motorola to coax 500MHz yields to a level that could support an Apple product. Apparently Motorola has finally succeeded, as Steve Jobs reannounced the 500MHz G4 at last week's Macworld-Tokyo conference. But even at 500MHz, PowerPC frequency is still behind Intel and AMD desktop processors by more than 40%

The G4's frequency deficit relative to Pentium III and Athlon is not a surprise, and it does not indicate a deficiency in Motorola's process. The real problem is with the short five-stage pipeline of the G4 processor, which gives good IPC but limits the almighty marketing parameter: frequency. The two additional pipeline stages on the new G4+ design will help, but the part will still be at a frequency disadvantage compared with the 12-stage pipeline of Coppermine, the 10-stage pipeline of Athlon, and the 20-stage pipeline of Willamette. Although the short pipeline may rationalize for readers of this publication the reason that the new G4 hasn't broken the gigahertz barrier, as Coppermine and Athlon have, it is an explanation that will fall on deaf ears among the PC consumers that Apple covets.

Motorola is not the only company that failed to put a gigahertz entry into the frequency race at ISSCC. HP, now focused on IA-64 for long-term salvation, mustered only a 600MHz chip, which we believe to be the PA-8600 (see *MPR 3/29/99-msb*, “PA-8600 Due in Early 2000”). Featurewise, this enormous 469mm^2 processor is nearly identical to its predecessor, the PA-8500, except that HP modified the LRU replacement algorithm of the on-chip single-level L1 data cache to improve its hit rate.

The PA-8600 is implemented in roughly the same IC process as the PA-8500, but the effective gate lengths were reduced to boost speed. At 600MHz, the PA-8600 is more than 35% faster than the currently shipping 440MHz PA-8500. HP says that 60% of this speedup is attributable to effective-channel-length (L_{eff}) reduction; another 25% came from critical speed-path tuning, another 5% from clock tuning, and the remaining 10% from system improvements. Demonstrating once again that frequency isn’t the only parameter that determines performance, however, HP reported that the PA-8600 would deliver an impressive 43 SPECint95 and 62 SPECfp95 at 552MHz. These performance figures will probably translate to baseline SPEC scores at 600MHz of approximately 42 SPECint95_base and 63 SPECfp95_base, putting HP 15% ahead of both the current SPECint leader, Pentium III Xeon-733, and SPECfp leader, Alpha 21264-700, but still 20–30% behind the 1-gigahertz 21264 that Compaq described at ISSCC.

Frequency Is Performance

Although gigahertz or near-gigahertz processors have been reported at ISSCC in previous years, this year there are many more, and each is much closer to reality. While no processor has yet broken the gigahertz barrier in production, that event will surely happen before the next ISSCC, probably several times. The likely winner of that race is not yet clear, but AMD with Athlon, Compaq with the 21264, and Intel with Coppermine and Willamette are all contenders.

The pursuit of high frequency has reached a frantic pace. There are two reasons: one is that frequency is a legitimate source of performance; the other is that the market judges the competitiveness of processors mostly on the basis of frequency—regardless of how well it reflects performance. To some extent, designers are rightfully focusing on the frequency vector, due to the painfully slow progress in instruction-level parallelism (ILP) through superscalar, out-of-order techniques. But some designers are now pressing on frequency even at the expense of ILP and without solid evidence the result will be higher performance. This market is clearly having an influence on designers’ engineering choices; whether or not it is a net-positive influence remains unclear.

A hot debate in the race toward higher frequency is the use of copper interconnects at the 0.18-micron level. On the nay side of the argument stands mighty Intel; on the opposite side stand most of the remaining semiconductor vendors. At this point, the evidence seems clear that copper is indeed valuable at the 0.18-micron level. By devoting as much time as it has to debunking this notion, Intel appears to be reacting like a polecat backed into a corner from which it cannot escape. Intel will no doubt survive the mistake—if indeed it is one—by throwing massive design resources at the wiring problem, but it should stop wasting time trying to convince the world it made the right choice. After all, the mistake is clearly not debilitating. The company has proved it can produce processors that are as fast as anyone’s, despite its use of aluminum wires.

While companies are continuing and even accelerating their drive toward higher frequency (if not higher performance), processors are rapidly outpacing software. The simple fact is—besides a few high-end server and workstation applications—today there are no high-volume applications that need gigahertz-class processors. The vast majority of PC applications work just fine with 600MHz processors, as evidenced by the increasing number of consumers who are opting for lower-cost PCs. If new applications don’t arrive soon to use the significantly higher performance processors that are in development, the growth of processor performance could be stifled long before Moore’s law plays out. In fact, compelling software, or lack thereof, could be the single largest problem facing processor vendors in the coming years. ◇