



## **Async chip design eases process shift**

By Uri Cummings (contributed article)  
January 14, 2003

We knew that our innovative asynchronous chip architecture could improve system efficiency by increasing performance and decreasing power consumption. But we had to show prospective customers that it could be done in a modern process technology. Our conclusion: Not only does asynchronous chip design work, but it also eliminates many of the traditional challenges of moving to state-of-the-art 150- and 130-nm process technologies.

The chip in question is the Vortex, which we believe is the largest and most complex asynchronous system-on-chip. It includes a novel nine-way superscalar processor with 32-kbyte instruction and data caches, a 10-Gbit Ethernet MAC, a quarter-terabit on-chip system interconnect and a DDR-SDRAM controller. It was the first chip designed by Fulcrum Microsystems for TSMC's 150-nm process technology. The device was functional on first silicon.

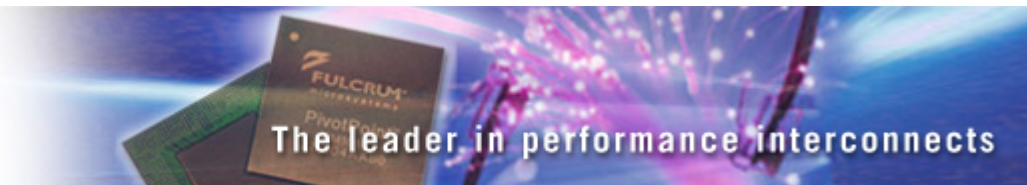
The Vortex is "clockless" in its design and uses Fulcrum's asynchronous pipelining mechanism to manage the data flow on the chip. In its simplest form, one circuit can pass a Data0 or Data1 bit to an adjoining circuit, which then sends an acknowledgment back to the original circuit. Once it receives the acknowledgment, the original circuit lowers its data signal and the receiving circuit, in turn, lowers its acknowledgment.

Fulcrum's implementation results in circuit designs that are naturally free of glitches and race conditions.

Because of the nascency of the asynchronous design style and the limited suite of commercial tools originally available to support the Fulcrum design flow, designing the Vortex in terms of design automation was like designing a semicustom chip. It took our team of 12 people about nine months to develop and optimize it. While the front-end design process was comparable in time and complexity to an equivalent synchronous ASIC, the clockless nature of the Vortex meant the schedule was not affected by any reengineering to solve the timing closure problems usually encountered with a synchronous chip the size of the Vortex and with its set of performance requirements.

The gate libraries and related tools used to develop the chip were built upon work done at California Institute of Technology and at Fulcrum, some of which date back to the company's first processor, which was designed in 1998 and used 600-nm technology. Fulcrum has developed a unique tool for generating gate library elements on the fly, sized specifically for each use. Effectively, this equates to a library with an unlimited number of custom-designed elements.

In the synchronous design environment, creating accurate transistor delay models is becoming increasingly difficult, and the models are subject to manufacturing variances, which are increasing as we move to processes with smaller and smaller features. As a result, synchronous gate libraries are relatively small, not as well optimized for each use and often padded with margin to address potential variances. We found with the Vortex



that our library was extensible and reusable in the 150-nm process with only minor tweaking to capture the performance improvements available in the more modern process.

With its delay-insensitive architecture, performance of circuits on the Vortex is separated from the functional correctness. Thus, an improperly sized gate leads to lower performance than perhaps expected, not an error condition (such as a race or a glitch), which is typically the case with synchronous designs.

### **Process variation**

On the Vortex silicon, we saw roughly a 10 percent in-die variation of power and frequency across the chip. This had no effect on our yields as the functionality of the Vortex' asynchronous circuits is not affected by variations in voltage or temperature. The chip's characterization results showed predictable performance across a broad range of voltage and temperature levels.

The 150-nm technology gave us a modest improvement in clock speed over chips fabricated in 180-nm technology. The Vortex operates at 450 MHz, which is roughly 10 percent higher than we saw in our initial F1 chip, which was fabricated in TSMC's 180-nm process. Interestingly, latency on the Vortex remained roughly the same as the 180-nm chips, since RC propagation delay remains roughly constant for a given die area as one moves to smaller feature sizes (aside from the future improvements introduced with copper and low-K).

But as the industry migrates to the modern processes and frequencies continue to increase, the relative propagation delay increases—that is, while it may take a few clock cycles to send a signal across a synchronous chip in 150-nm technology, it will take several clock cycles to do the same thing in 130-nm technology at a proportionally higher clock rate. Synchronous designs will increase in complexity as they attempt to overcome this hurdle. Fulcrum's delay-insensitive design style will not be affected by proportionally increasing propagation delays.

Also, the smaller feature size of transistors in 150 nm decreased the latency through the Vortex' domino logic. Because domino logic generally eliminates "crowbar" current and uses primarily N-type transistors for computation, it has very fast forward latency through the chip.

At 13.5 million transistors, Vortex would be considered a moderately large and complex chip for 150-nm technology. Complexity was affected slightly by Fulcrum's use of additional wires for the handshaking protocol. In our first processor—at 600 nm—this was a design challenge as the extra wires consumed a lot of real estate on the die. However, with several metal layers and shrinking feature sizes in TSMC's modern processes, accommodating the extra wires has become a very manageable issue. We expect this to be even less of a consideration as we move forward to a 130-nm process, with copper interconnect replacing aluminum.

Another advantage of the Fulcrum delay-insensitive architecture is alleviation of any power profile problems the Vortex might have had in 150 nm. A byproduct of asynchronous technology is clock gating down to the individual circuit level as each circuit naturally consumes no power (except, perhaps, for nominal leakage current) when not in use, which is combined with as much as a 30 percent improvement in power consumption due to the lack of a clock and its associated distribution tree and latches.



The Vortex chip gave us proof that our asynchronous architecture could be applied to today's most cutting-edge process technology. Soon we'll have characterization results for a next-generation version of the crossbar that we expect will have more than 1.2 Tbits/second of nonblocking capacity-in a 2-mm<sup>2</sup> footprint.

As designed, the next-generation switch will operate at more than 1 GHz, drawing fewer than 2 W during peak performance. The performance of this chip will push the boundaries of the 130-nm process, but even then we anticipate that because we don't distribute a clock we eliminate many issues of interconnection between blocks and noise.

Our experience in developing the crossbar switch for the 130-nm process was the same as with the Vortex. Although we needed to do some engineering work to accommodate new design rules for the new technology, the asynchronous architecture of the chip insulated the chip dramatically from the problems that are common with adopting a new process technology.