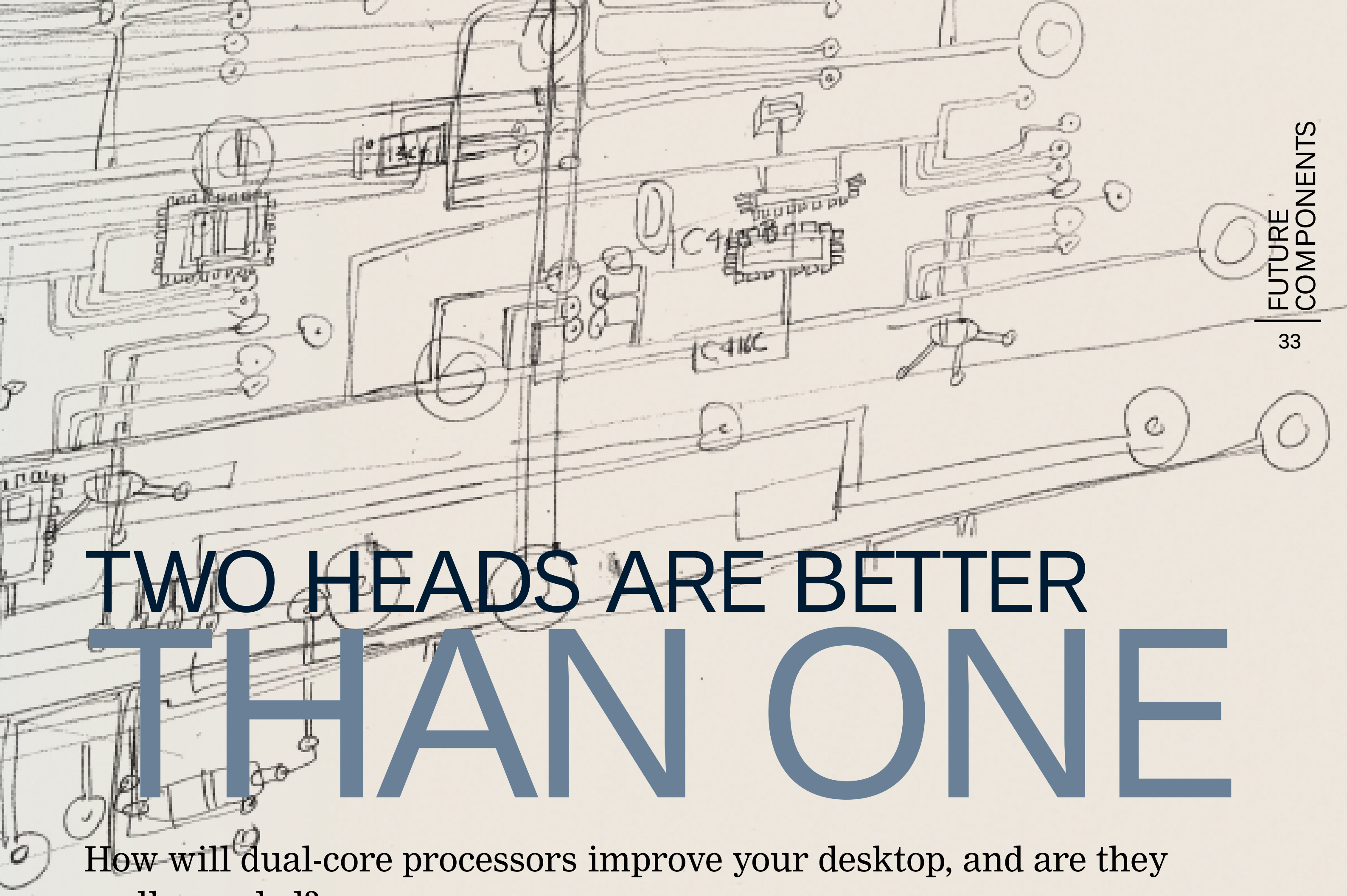


TWO HEADS ARE BETTER THAN ONE

How will dual-core processors improve your desktop, and are they really needed?

By William Knight

TRADITIONAL APPROACHES to processor design are at a watershed. Shrinking die sizes and increasing bond-pad densities are making it harder to establish contact with the outside world, while escalating clock speeds are boosting heat dissipation problems to the point where processors are in danger of melting a hole in the motherboard and burning the carpet. On the architectural front, sophisticated pipelining techniques – splitting instructions into sub-components for parallel execution – seem to have reached an optimum balance between improving the best case and doing minimum harm in the worst case.

In response to this technology impasse, processor manufacturers are turning to dual-core designs, combining two independent processors and their respective caches and controllers onto a single silicon chip; thereby neatly increasing performance without changing architectures or increasing clock speed. Intel and AMD both offer dual-core processors. Intel was first out of the blocks in April 2005 with the Pentium D (and later, Pentium D Extreme edition), followed by AMD, offering dual-core versions of its popular Athlon and Opteron processors.

According to Dave Everitt, European product and platforms manager for AMD, the industry has been planning dual-core processing for some time. AMD's Hammer architecture – the basis for AMD's Athlon64 (desktop) and the Opteron (server) chips – first announced in 1999 at the Microprocessor forum, was designed to be multi-core from its inception.

CLOCK SPEED IS NOT PERFORMANCE

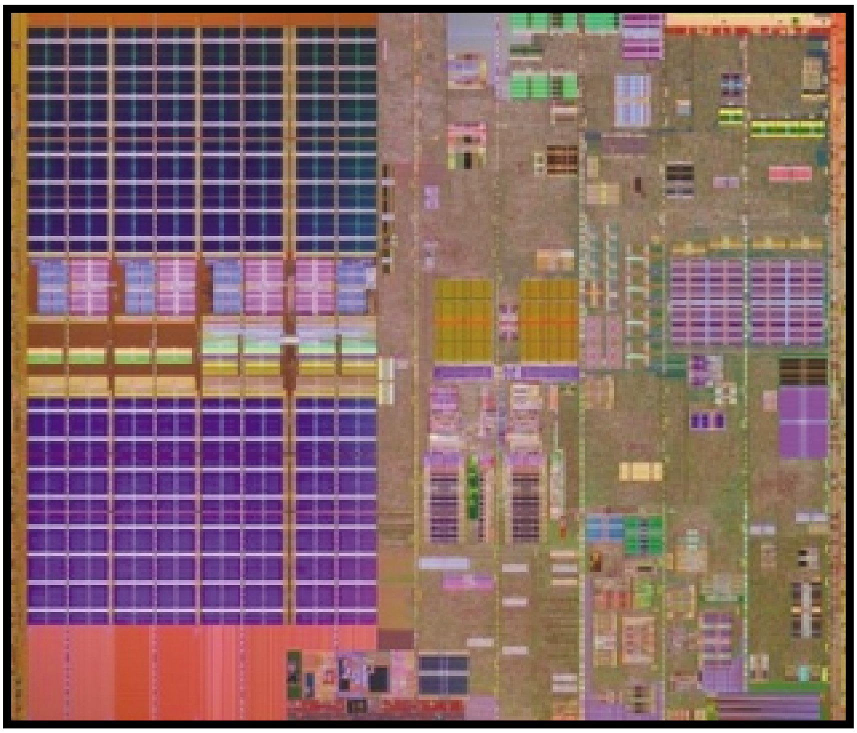
The journey to dual-core has not been straightforward, with the advantages of this approach only becoming clear after the full exploration of single-core design capabilities.

For a start, building multi-core processors using a familiar architecture presents fewer development risks because the architecture is stable and known. Attempting to get the same boost using a single-core approach will almost certainly require considerable design changes and operating system updates.

Furthermore, succeeding generations of processors have generally involved an increase in clock speed. However, doubling the clock speed does not double performance. Alan Priestley, strategic enterprise marketing manager at Intel, explains, "We don't always get a significant performance increase. The faster the frequency increases, the less you get in terms of performance gains without some architectural change on the processor."

Additionally, a dual-core processor will run far cooler than a single core design of equivalent performance running at double the clock speed. For example, on a chip produced using AMD's 90nm process, power consumption increases by 60% with every 400MHz rise in clock speed, so that a ➜

Illustration: Daniel Mackie

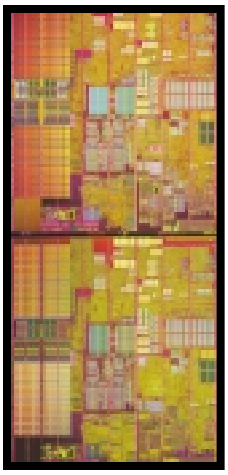


doubling in the current maximum clock frequency would simply melt the processor. But the dual-core approach means you can get a significant boost in performance without the need to run at ruinous clock rates.

Pipelining – effectively increasing parallelism at the instruction level – has been a mainstay of processor design, with each sub-component of an instruction assigned to a separate execution unit. The AMD Hammer architecture has nine such execution units, which, potentially at least, increases the performance of the processor nine-fold. Everitt explains how pipelining aids performance by the analogy of a walk with his young daughter. She takes two steps to his one, her clock speed is twice as fast, but they get to the sweet shop at the same time because he does double the work with each step. "She is obviously running at a higher frequency than I am, but we are still getting the same job done," he says.

In practice, keeping all these pipelined execution units fully employed is a real challenge, dependent on the ability to identify instructions that can be executed out of their natural sequence. As the depth of the pipeline is increased (potentially increasing performance) the risk of an inappropriate out-of-sequence instruction execution increases. The need to repeat such executions presents an inescapable constraint on the benefits of pipelining.

Intel uses fewer execution units but instead employs hyper-threading, or simultaneous multithreading, in which two programs, or threads, may be loaded into one core. Hyper-threading essentially fools the operating system into thinking it's connected to two processors, so that the threads run on two 'logical' processors, created within a single physical processor. To achieve the same effect without hyper-threading, the operating system must swap each program in and out of memory, sharing execution time on a schedule or priority basis.



Top: Intel's single-core Pentium processor
Above: Dual-core, Extreme Edition Pentium

Swapping is clearly redundant effort, but splitting the core by hyper-threading only helps when two programs are being run at the same time – a single program in a dedicated core would actually be better off without hyper-threading. Critically for Intel, the desktop environment is replete with applications vying for time, and users are demanding better performance from background tasks.

DESKTOPS NEED MULTI-CORE

In 1995 the Internet was in its infancy, but now we are running video downloads, broadband, real-time virus scanners, intrusion detection systems, Ethernet connections, DVD burners and so on, and all these applications, many of them resident in memory, take their toll on performance.

Faced with these demands on processing resources, Priestly has no hesitation in dismissing the idea that dual-core represents technology overkill – the equivalent of a four-wheel drive vehicle on the school run. "We've not got to a point yet where the software community can't use the capabilities we are building," he insists. "We are opening up new possibilities in terms of functionality, user-interface, usability, absolute system capability. We've not capped out yet."

"We are still using keyboards. It would be much easier if I could talk to it [the computer], but to talk to it and get damned good speech recognition requires a hell of a lot of processing power."

However, maximum performance is not all down to hardware, and to take full advantage of multiple cores and Intel's hyper-threading, software must be designed with simultaneous execution in mind.

Conventionally, programs are loaded into memory as a long line of instructions. The processor takes the first instruction, executes it, then gets the next instruction and so on until it reaches the end of the program. This is the single-threaded approach. But if, using clever design, the application can be divided into multiple threads – for example a thread that fetches email and a thread that renders the screen image – the threads can be executed simultaneously, providing a distinct performance boost.

For a processor operating in a given power envelope, more cores will out-perform fewer cores for multi-threaded applications, for multi-tasking response times and for transaction processing. But single cores will be better for single-threaded, performance-sensitive applications.

Designing multi-threaded software can be very involved, and while hardware has been unable to support multiple threads – except via the crude time-slicing of the operating system – there has been little demand for developers to supply multi-threaded applications to consumers. Now dual-core is a reality, updated software is trickling onto the market. Adobe Photoshop is multi-threaded, and the first multi-threaded games will be available shortly.

Even if applications are not dual-core enabled, it will still be possible to get a boost by running independent programs simultaneously; for example by doing a spell check while burning a DVD in the background.

DUAL-CORE AS A SUBSET OF MULTI-CORE

Dual-core is only the start. As silicon real-estate is released through technical advances, expect a rapid move to multi-core processors, fuelled by our ravenous appetite for multimedia. Currently, AMD and Intel are working at 90nm feature sizes, but 65nm will go into production shortly and 45nm is only a few years away. At 130nm, a single-core version of the AMD Opteron has a die size of 193mm^2, while at 90nm the dual-core Opteron is only slightly larger at 199mm^2.

Processing advances are, undoubtedly, one of the drivers behind the move to multi-core. "The geometries are beginning to come down, the wafer sizes are going up, the die sizes are staying [roughly] the same, so the amount of workable silicon area is getting bigger. And the question is, what do you put in that workable silicon area?" asks Everitt. He thinks one answer is obvious: more cores.

THE AMOUNT OF WORKABLE SILICON AREA IS GETTING BIGGER. THE QUESTION IS, WHAT DO YOU PUT IN THAT WORKABLE SILICON AREA? – *DAVE EVERITT, EUROPEAN PRODUCT AND PLATFORMS MANAGER, AMD*

The Cell multiprocessor, released by IBM, Sony and Toshiba, is a direct response to the boom in multimedia: streaming video, life-time memories and an explosion in Internet voice traffic. With roughly the same dimensions as dual-core chips from Intel and AMD, and with a similar transistor count, (234 million), The Cell contains a 64-bit PowerPC core and eight, so-called, 'synergistic processing elements' (SPEs) on one 221mm^2 die.

The SPEs are powerful processors focused on churning through single-precision and double-precision mathematical calculations. Each SPE gets its orders from the central Power PC which handles the scheduling and parcelling of data, leaving the SPEs free to concentrate on pure number crunching.

The gaming world is humming with expectation as the Cell goes into Sony's PlayStation 3, but IBM thinks Cell's innovative design is a perfect fit in TVs, set-top boxes, and a host of multimedia applications, from entertainment to medicine and defence. Evidence for this last viewpoint is provided by the recent announcement of Mercury Computer Systems as IBM's first development partner for the Cell. Mercury Computer is expert in the real-time processing of radar, sonar and intelligence data, indicating that the Cell is slated for a wide range of applications, all centred on data analysis and interpretation.

It can't be long before Cells start to appear on desktop PCs, but currently, as with dual-core, the challenge is building software to make the most of the new-wave parallelism, and to create multi-threaded applications that will deliver real performance benefits. However, once software coders become armed with the appropriate development tools and learn the tricks of parallel programming, multi-core will become the norm.

"The software landscape has changed," says Priestley, "We've got operating systems that can support running on multiple CPUs (or multiple execution units), and now we are starting to get applications that can do that as well. And that then facilitates the market place for us to deploy dual-core CPUs". ■