By Brian Dipert, Technical Editor

Illustration by Daniel Guidera

YOU WANT CHIPS WITH RICHER FEATURES AND CORRESPONDINGLY HIGHER PIN COUNTS. BUT, DUE TO EVER-SHRINKING SYSTEM FORM FACTORS, YOU ALSO WANT THOSE CHIPS TO BE THINNER AND TO CONSUME LESS SYSTEM-BOARD AREA. WHAT'S A SEMICONDUCTOR SUPPLIER TO DO?

Silicon contends with STUFFED AND SHRINKING PACKAGES

At a glance **50**
Future fashions **52**
Standards shortfall and a subsystem perspective **54**
For more information **58**

"More complex" and "smaller" are two common requests that system designers ask of chip vendors, often in the same breath. In the very next breath, invariably, they ask the vendors for devices that are ever faster, lower power, and lower priced, too. Packaging breakthroughs are key to enabling vendors to continually meet or exceed your seemingly divergent needs.

Looking first at the big picture, you can group the primary functions of a chip's package into three main categories:

- the traffic cop, which routes power, ground, and other inputs to the die or dice and routes die or dice outputs to the rest of the system as quickly and with as little added distortion as possible;
- the firefighter, which draws performance-sapping and reliability-limiting heat away from the die or dice to be absorbed by the surrounding environment; and
- the soldier, which protects the die or dice from mechanical stresses and other environmental effects, while creating no stresses of its own.

Innovations in packaging technology often receive less publicity than advancements in the silicon inside them. But evolutionary and revolutionary packaging achievements are, if anything, more impressive than the now-predictable Moore's Law trends. We live in an era in which many analysts are forecasting that packaging costs will soon dominate the total cost of both low- and high-end ICs (**Figure 1**). As a result, adequate packaging research and development, and the breakthroughs resulting from that investment are critical to

continued industry advancement.

How can you and your semiconductor suppliers work together to resolve your multifunction and multiproduct needs? The traditional scenario involves multiple suppliers delivering multiple chips, each in the smallest package possible, which you then attach to and interconnect on the system board. On the other end of the spectrum is the highly hyped system on chip, which combines all necessary functions on a single piece of silicon. Between these two extremes is the multichip module, combining several unpackaged die or already-packaged devices in a horizontal miniature pc board or vertical stack under a common package lid. These options all have trade-offs. Engineers have different system needs and priorities, and they invariably come to different conclusions about which option is best for them.

AT A GLANCE

▷ Continued packaging innovation is necessary to keep the momentum of Moore's Law.

▷ As chips grow more complex in their functions, array packaging, especially bump-based, will likely grow in popularity.

▷ Package options that look the same from the outside may have different internal architectures.

▷ Multichip modules make sense for many reasons, but cost savings at the device level may not be one of them.

LEGS AND BUMPS

Leaded (legged) packages, such as DIPs, PLCCs, SOPs, and QFPs represent the lion's share of all packages used today. Also known as "boundary" packages (a category that also includes leadless packages), their connections to the outside world align along one or multiple edges of the package, and they come in both through-hole and surface-mount connection options. Surface-mount packages are becoming increasingly popular, except possibly in prototyping environments, in which easy replacement of chips is desirable, and in manufacturing environments that value the just-in-time insertion of a chip before shipping the system to a customer. Sockets are options in these cases. However, they increase system cost, and their added impedance may negatively affect performance and signal integrity.

Speaking of impedance, it's one reason, though not the primary one, that array-packaging contenders, such as BGAs and PGAs are rapidly growing in popularity (**Figure 2**). Ball, or bump, connections offer lower impedance than leads. More important, as device pin counts increase, array packages *usually* offer more space-efficient board footprints than do boundary packages.

The "usually" qualifier is important for several reasons. It's common today to find QFPs with lead-to-lead spacing, or pitch, measuring 0.05 mm, because it's often unnecessary to route signal traces between the QFP leads. Instead, you can bring signals to the top of the board using vias within the bare cavity that the pinout boundary defines and then fan out the signals to the leads. Conversely, most array packages have no cavity; some have a small one. Regardless, you almost always need to route traces between bumps to reach the interior of the array. As the array grows, the number of signals to route increases. Fine-pitch traces, multilayer boards, and buried vias all significantly increase board cost.

Today's high-pin-count chips, therefore, rarely employ bump pitches less than 1.2 or 1 mm. Lower-pin-count and, therefore, smaller-array parts, such as flash memories, might use a 0.8-mm pitch, and you can find even 0.5-mm pitch packages for use in extremely space-sensitive applications. Low-impedance BGA sockets, often employing elastomeric connections, come from companies such as Ironwood Electronics. As an alternative to BGAs, companies such as Advanced Interconnect Technologies are advocating next-generation boundary packages, such as the no-lead

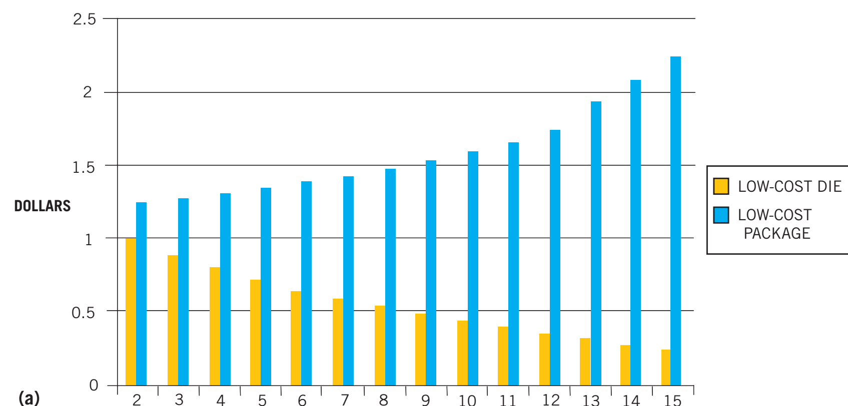
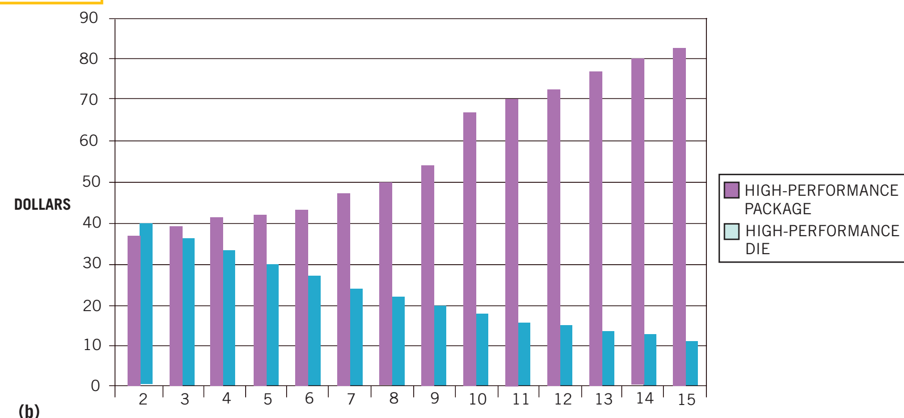


Figure 1



Industry projections suggest that packaging costs soon will outweigh silicon-die costs for both low- (a) and high-complexity (b) devices (courtesy *IEEE International Technology Roadmap for Semiconductors*).

VFQFP and thinner WFQFP. The company claims that these packages deliver BGA-like board footprint areas for all but the highest pin counts and much lower impedance than do traditional leaded packages. VFQFP and WFQFP come in both wire-bonded (**Figure 3**) and flip-chip-based versions.

The PGA is the odd man out in the boundary-versus-array packaging debate. Like a through-hole boundary package, you can easily socket it. (This feature is important in situations in which, for example, a PC manufacturer wants to support numerous processors and, therefore, system-performance speeds with a single board design.) Like array packages, PGAs are board-space-efficient. They ship in huge volumes, driven by PCs, but, due to their costly requirement for either a socket or a fine-pitch through-hole pc board, they're largely unpopular outside that application.

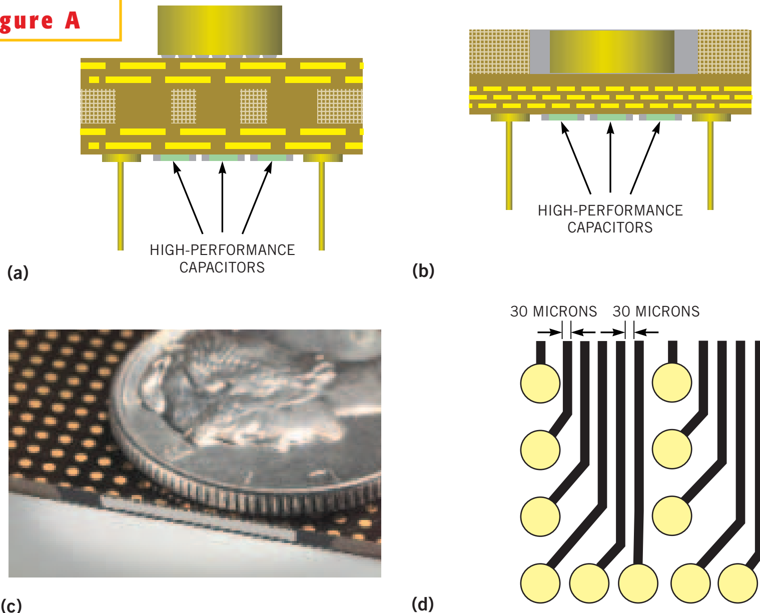
The terminology you encounter in investigating array packages might confuse you. Specifically, the acronyms BGA and CSP often seem to refer to packages that look identical. Historically, a CSP was a package that, using common industry conventions, was no more than 20% larger than the die it housed. BGA technology advancements over the past few years

FUTURE FASHIONS

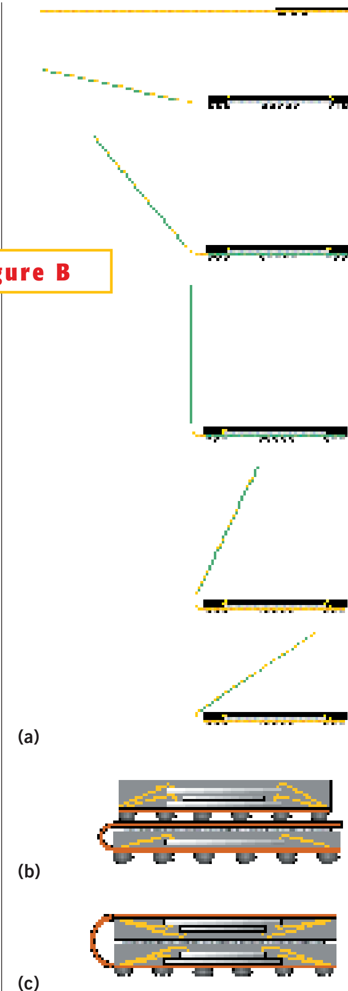
Reducing transistor sizes at Moore's Law-predicted rates is by itself insufficient to ensure ever-faster chips. Vendors also need to reduce the propagation delay of signals passing from one transistor to another—for example, by switching from aluminum to copper interconnect. And they'll eventually need to accelerate the device substrate by migrating away from silicon oxide to a material with a lower dielectric constant (known as the k factor).

USG (undoped silicon glass) has a dielectric constant of approximately 4.5. The dielectric constant of FSG (fluorosilicated glass) is about 3.7, and low-k dielectric materials currently being researched are at 2.7 and less. Air, with a k of 1, is, at least from an electrical-transmission standpoint, the ideal substrate material! However, as air exemplifies, materials with decreasing dielectric constants also have weaker mechanical structures, putting more of the overall structural burden on the surrounding package. This phenomenon is part of the reason that the historical package-versus-die demarcation line will in the future become increasingly blurred.

Look, for example, at Intel's BBUL (bumpless-buildup-layer) package, scheduled for volume production before the end of this decade. Instead of using a flip-chip-based die sitting on top of the packaging substrate, BBUL embeds the die within the package boundary (**Figure A**). The approach has numerous benefits beyond structural stability: no more bumps (BBUL instead directly and speedily connects the die to vias embedded within the package substrate), a much closer proximity between the die and its bypass and filter capacitors, and a much thinner resultant package. The lack of



Whereas conventional packages locate the die on top of the substrate (a), BBUL (bumpless-buildup-layer) packaging surrounds the die (b). Low height is one of the resulting advantages (c), along with more flexible die-to-substrate interconnection (d) (courtesy Intel).



Folded, stacked interconnect (a) creates quick time-to-market combinations of single-die and multidie packaged chips in logic-plus-memory (b) and memory-only (c) configurations (courtesy Intel).

have made this 20% rule the norm rather than the exception, so more and more people are using the two acronyms interchangeably.

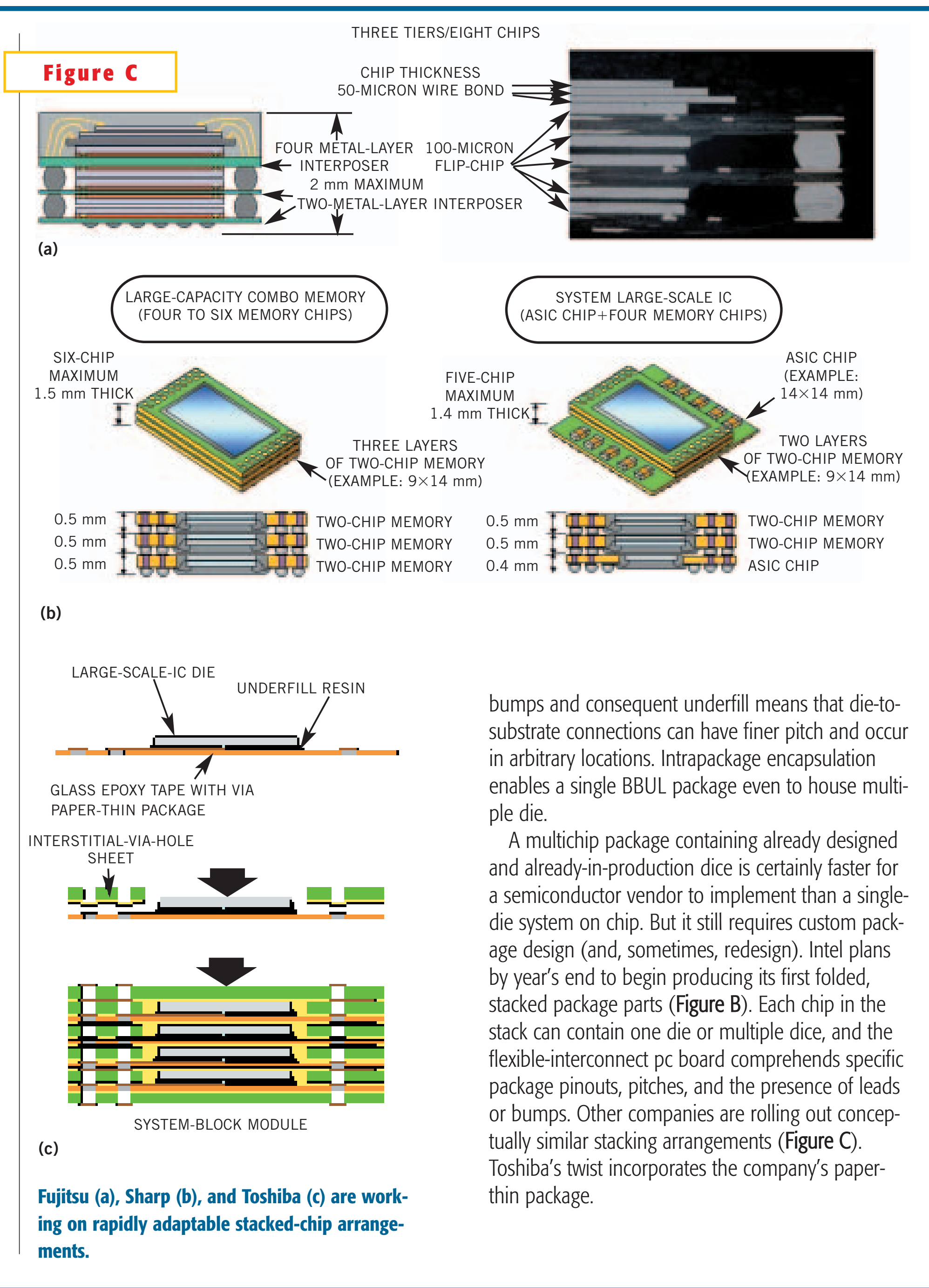
Given the increasingly svelte array and boundary packages now housing die, what place do bare die have in the packaging smorgasbord? In the past, manufacturers used them in the most space-constrained applications, and they are still key ingredients in multidie chips. But they're difficult to handle in a manufacturing environment. The required bond wires or balls and the encapsulating overmold increase system cost and negate subsequent board-rework capability. And, suppliers' abilities to test bare die for speed and reliability are limited.

LOOKING UNDER THE LID

Package alternatives may look similar from the outside. Dissect them, though, and their differences become more apparent. Two primary factors—heat dissipation and signal integrity—determine the orientation of the die relative to the substrate and the interconnection between the die and substrate. Looking first at the conventional BGA or CSP, you see that the material lying between the die and the outside world is the overmolding of ceramic or plastic (**Figure 4a**). You also notice that bond wires interconnect the die and the substrate. A bond-wire variant, TAB (tape-automated bonding), patterns the die-to-lead-frame interconnections on a multilayer polymer tape.

Now, compare the conventional BGA with a cavity-down BGA (**Figure 4b**). The die-to-substrate interconnect scheme is still the bond wire. But, with the die now mounted upside-down, the vendor can directly attach a metallic heat spreader or heat sink to its backside.

Finally, take a look at the flip-chip BGA (**Figure 4c**). Again, the die is upside-down for direct attachment of a heat spreader. But this time, the interconnection between die and substrate consists of bumps, not bond wires. Why? Perhaps this estimate from Vincent Wang, PhD, Altera's senior director of packaging, will provide a clue. Wang estimates that, whereas the typical bond wire has an inductance of 4 nH, a flip-chip bump's inductance is 0.4 nH. Reduced inductance means reduced signal distortion and increased propagation speed, along with rock-solid power and ground connections and robust heat transfer away from the die. And chip designers can put bump connections in optimal die locations, instead of, by necessity, placing them all on the die boundary. Intel, for example, has converted nearly 100% of its PC-mi-



Fujitsu (a), Sharp (b), and Toshiba (c) are working on rapidly adaptable stacked-chip arrangements.

bumps and consequent underfill means that die-to-substrate connections can have finer pitch and occur in arbitrary locations. Intrapackage encapsulation enables a single BBUL package even to house multiple die.

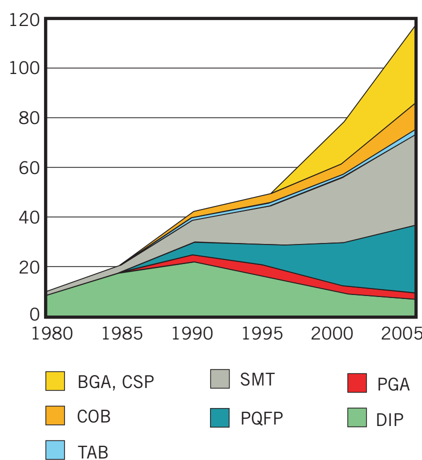
A multichip package containing already designed and already-in-production dice is certainly faster for a semiconductor vendor to implement than a single-die system on chip. But it still requires custom package design (and, sometimes, redesign). Intel plans by year's end to begin producing its first folded, stacked package parts (**Figure B**). Each chip in the stack can contain one die or multiple dice, and the flexible-interconnect pc board comprehends specific package pinouts, pitches, and the presence of leads or bumps. Other companies are rolling out conceptually similar stacking arrangements (**Figure C**). Toshiba's twist incorporates the company's paper-thin package.

croprocessor production to flip-chip-packaged devices.

Higher performance, alas, doesn't come for free. The bump-attachment process increases die cost. The required ultrafine bump-contact pitch has significantly higher package-substrate cost than a bond-wire-mated alternative (see **sidebar** "Future fashions"). And, for added stability, the gaps in the matrix between bumps become filled with an undermold material that hardens under microwave exposure, adding more cost. For these reasons, Altera's Stratix FPGAs will come in a mix of packages: bond-wire-based for low cost in low-chip-speed bins and flip-chip-based for the fastest, most expensive chip-speed bins. Note that bond-wire and flip-chip interconnection options are generic to both array and boundary packaging.

Another key function of the undermold is to buffer the differing coefficient of expansion and contraction (t_{CE}) of the die and the substrate. Bond wires also act as mismatch buffers. Tessera's μBGA package dispenses with the substrate, relying instead on a flexible elastomer material. The downside of Tessera's approach, though, is that it's closely tailored to the die size and layout; a chip redesign or process-lithography shrink invariably forces changes in the package dimensions, along with the bump-matrix pinout and pitch. More conventional substrates, which you can conceptually think of as mini pc boards, hide these die-level changes from the outside world.

Figure 2



Projections hint at a gradual shift from boundary to array packaging in coming years (courtesy Altera and Amkor Technology).

Mismatches in t_{CE} between the substrate and the system pc board also require careful consideration. One of the factors driving the switch to organic substrate materials is that their t_{CE} has closer equivalence to FR4 and other common pc-board raw materials. When considering temperature effects, keep in mind that both the absolute minimum and maximum temperatures and the rate of temperature cycling between those extremes are important. The system can experience rapid temperature changes during both normal operation and manufacturing (during wave soldering, for example).

And, speaking of wave soldering, it's a common environment to experience another critical packaging-failure mechanism, this one related to moisture. Have you ever wondered why chips come with stringent specifications on how long you can keep them out of their desiccant packaging before soldering them to the board for or how long you need to bake them before soldering if you've exceeded this time threshold? Packages absorb atmospheric moisture; think of them as miniature sponges. Subject them to extreme heat, and, if they contain more moisture than can escape as steam, they'll pop right off the board—a phenomenon not surprisingly known as "popcorning."

STANDARDS SHORTFALL AND A SUBSYSTEM PERSPECTIVE

Look at today's SRAM-plus-flash-memory stacked BGA chips, and you'll find two main pinout "camps." Ironically, they both employ a 0.8-mm bump pitch. And (sigh) they're *both* JEDEC standards. Before the kickoff of the official standards process, two incompatible approaches emerged—one based on the so-called "Intel and Sharp" flash-memory pinout and the other based on the "AMD and Fujitsu" pinout. JEDEC found, when it finally got around to discussing the topic, that both options were entrenched in the marketplace and that compromise between them was impossible, so they rubber-stamped both. Intel and Sharp advocates claim that AMD and Fujitsu chips require either custom SRAMs or complex pin-scrambling substrates. AMD and Fujitsu fans predictably dispute these claims.

Memory modules (DIMMs, RIMMs, and SIMMs) are a common means of circumventing vendor-to-vendor chip package and pinout incompatibilities. They also provide density flexibility. Avido Systems takes the memory module to the next level of integration with its FlexCard family (**Figure A**). Employing industry-standard 144- and 200-pin small-outline DIMM connectors, FlexCards can contain 16 to 256 Mbytes of 32- or 36-bit SDRAM. They also can include 256 kbytes to 64 Mbytes of parallel interface (16- or 32-bit) flash memory, as much as 8 Mbits of SPI flash memory, and as much as 256 kbits of I^2C interface serial EEPROM. Serial EEPROM stores a specific device description identifier.

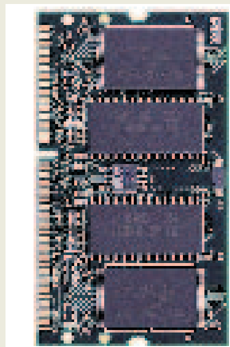


Figure A The FlexCard combines serial- and parallel-interface nonvolatile and volatile memories in a range of densities, all on one module (courtesy Avido).

HUNGRY FOR A SHORT STACK?

Some systems are so space-constrained that simply shrinking the footprint and height of each chip's package is insufficient. At this point, you need to consider combo packaging, which unites multiple die or already-packaged chips under a common encapsulation. Or, more

pragmatically, an MCM (multichip module) may be appealing to you, simply because it reduces the number of chip vendors from which you need to obtain products.

The MCM concept has been in use for a long time; IBM mainframes several decades ago employed TCM (thermal-conduction-module) packaging. Until recently, however, its high cost has limited its widespread adoption. Even today, potential chip-cost savings should *not* be a primary reason to consider an MCM, although companies such as Sharp claim that they've perfected the MCM manufacturing process to the point of near cost parity with the bag-of-chips alternative.

System-cost savings *are* possible, however. By shifting some of the chip-to-chip interconnect away from the system pc board and onto the package, you might be able to reduce the complexity and cost of the system pc board. And, as cellular-phone manufacturers have discovered, even if a stacked flash-memory-plus-SRAM MCM is more expensive than a two-chip substitute, you can sometimes more than make up the difference by charging a higher price for the much smaller resultant system.

MCMs' biggest success story to date has come with advanced cellular phones, in the form of a linear (DiNOR, NOR) flash-memory-plus-SRAM combo (see **sidebar** "Standards shortfall and a subsystem perspective"). As the required amounts of code-, data-, and file-storage memory increase with the phone's transformation into a multifunction device, manufacturers are adding serial flash memory (AND, NAND) and DRAM to the mix, resulting in three- and four-die chips. Fujitsu claims that it can even provide eight-die stacks for specialty applications. Between-die spacers resolve differences in die size and orientation.

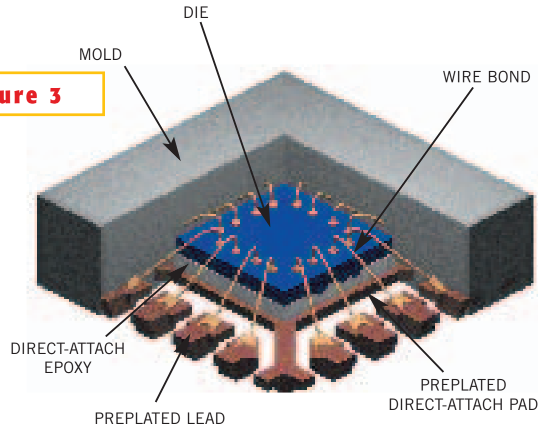
Ordering an MCM might be a simple matter for you, but it's an integration headache for your supplier, especially if it doesn't manufacture all of the dice inside the package. First off, to avoid creating an unacceptably tall package, the manufacturer must back-grind each die to reduce its thickness. It's also difficult to impossible to fully test a device in bare-die form; speed is a function not only of the die itself but also of the package interconnect and lead frame, and infant-mortality testing to screen out latent silicon-oxide defects must occur at high temperatures (a process called "burn-in"). Burning in an MCM to test one die might adversely affect another die in the module. And if one of the die in an MCM fails testing, the manufacturer ends up with a much more expensive paperweight than with a traditional single-die-packaged part.

If the vendor creating the MCM lacks in-house expertise in testing a given die (such as with a flash-memory supplier that obtains SRAM from other companies), procuring sufficient ongoing process and product knowledge to ensure adequate testing coverage can be a legal and logistical negotiation nightmare. Dice from different manufacturers have different dimensions and require requisite packaging adjustments. And every time a manufacturer redesigns one of those die, package readjustments are also necessary. MCM strategies to reduce discards include employing dice manufactured on trailing-edge processes (where oxide defects are well- understood and minimized) and not pushing speed (to maximize ac-testing yield).

Fortunately, all of the vendors this article covers claim that multimemory MCMs have no thermal issues that might prevent you from, for example, reading from the SRAM while programming or erasing the flash memory. Thermal concerns become more significant, however, when you consider a logic-plus-memory MCM. Sharp, for example, currently ships multidie combinations of its BlueStreak ARM-based CPU and various memories, both to its internal systems divisions (think camcorders and PDAs) and to key external customers.

In the past, notebook-computer manufacturers ensured graphics-subsystem memory and manufacturer flexibility by employing an AGP- or PCI-connector-

Figure 3



Some vendors claim that fine-pitch, small-footprint, leadless boundary packages, such as VFQFP and WFQFP, are lower cost alternatives with equivalent performance to array packages, especially at moderate pin counts (courtesy Advanced Interconnect Technologies).

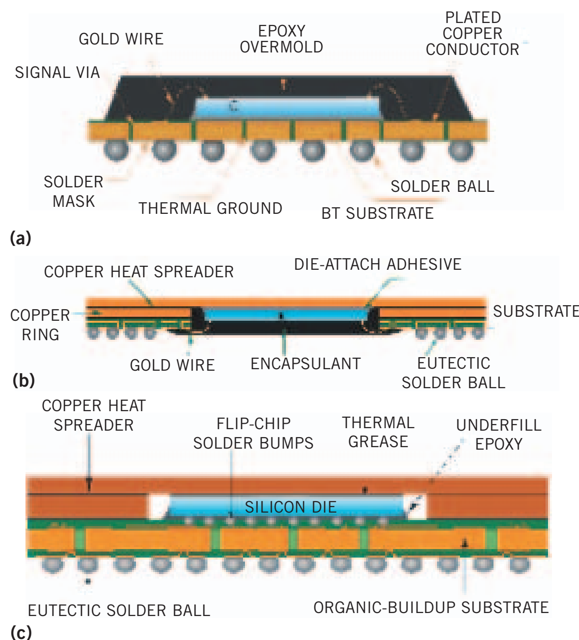


Figure 4

Conventional BGA (a), cavity-down BGA (b) and flip-chip BGA (c) packages deliver different combinations of cost, heat-dissipation, and performance characteristics (courtesy Xilinx).

based graphics minimodule. Today's thin and light computers have insufficient room for such modules, though. Nowadays, mobile graphics accelerators from companies such as ATI Technologies and Nvidia commonly combine logic and DRAM chips within a single MCM. Historically, graphics suppliers used bare die, but to ensure more complete test coverage, they're now integrating partners' BGA-packaged memories. Partially to circumvent heat concerns, these companies horizontally orient the multiple chips inside the MCM, although Nvidia combines four memory chips and a graphics die in a pseudovertical-stack arrangement. A horizontal orientation, however, doesn't save board space, and lack of board space is often a primary motivator for an MCM in the first place.

More generally, logic-plus-memory MCMs tend to use lower voltage and slower and, therefore, less-heat-generating, logic than you would commonly find use in discrete chips. One other testing issue that affects your system-board-manufacturing throughput is whether you can independently access multiple dice inside the MCM (and thereby be able to test them in parallel) or whether you need to access one or several of the die "through" another die, resulting in a slower serial-testing sequence.□

References

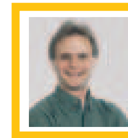
1. Schweber, Bill, "Lead-free mandate plumbs new design challenges," *EDN*, April 18, 2002, pg 55.

2. Israelsohn, Joshua, "Packaging solves 'the last centimeter'," *EDN*, May 24, 2001, pg 73.

Acknowledgments

I'm especially grateful to Vincent Wang, PhD, Altera's senior director of packaging, for identifying and explaining packaging trends within both his company and the industry at large.

Author's biography



Technical Editor Brian Dipert's final project at his previous employer involved coordinating the worldwide introduction of a μBGA package for flash memory...and he still has flashbacks. What a long, strange trip that was. You can reach Brian at 1-916-454-5242, fax 1-916-454-5101, bdipert@pacbell.net, and www.bdipert.com.

FOR MORE INFORMATION...

For more information on products such as those discussed in this article, go to www.ednmag.com/info and enter the reader service number. When you contact any of the following manufacturers directly, please let them know you read about their products in *EDN*.

Advanced Interconnect Technologies (AIT)
1-925-426-3100
www.aitsales.com
Enter No. 314

Altera
1-408-544-7000
www.altera.com
Enter No. 315

Avido Systems
1-908-276-9600
www.avido.com
Enter No. 316

Fujitsu
1-408-922-9000
www.fujitsumicro.com
Enter No. 317

Intel
1-408-765-8080
www.intel.com
Enter No. 318

Ironwood Electronics
1-651-452-8100
www.ironwoodelectronics.com
Enter No. 319

Nvidia
1-408-486-2000
www.nvidia.com
Enter No. 320

Sharp
1-360-834-2500
www.sharpsma.com
Enter No. 321

Tessera Technologies
1-408-894-0700
www.tessera.com
Enter No. 322

Toshiba
1-949-455-2000
www.toshiba.com
Enter No. 323

OTHER COMPANIES MENTIONED IN THIS ARTICLE

Advanced Micro Devices (AMD)
www.amd.com

Amkor Technology
www.amkor.com

ATI Technologies
www.atitech.com

Institute of of Electrical and Electronics Engineers (IEEE)
www.ieee.org

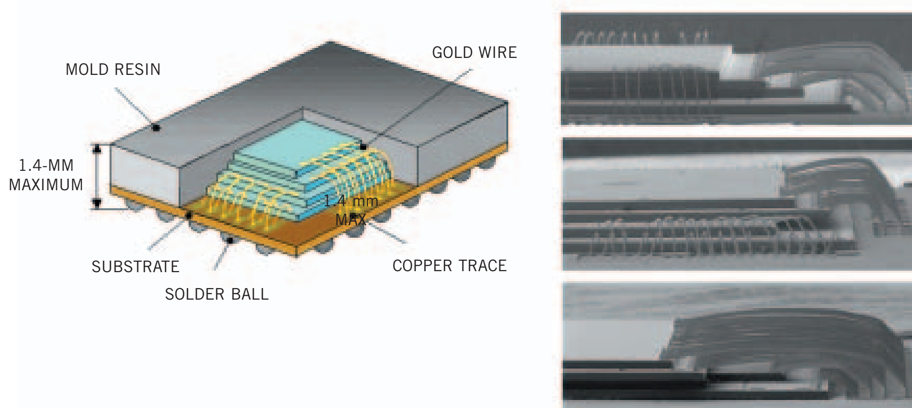
IBM
www.ibm.com

Joint Electron Device Engineering Council (JEDEC)
www.jedec.org

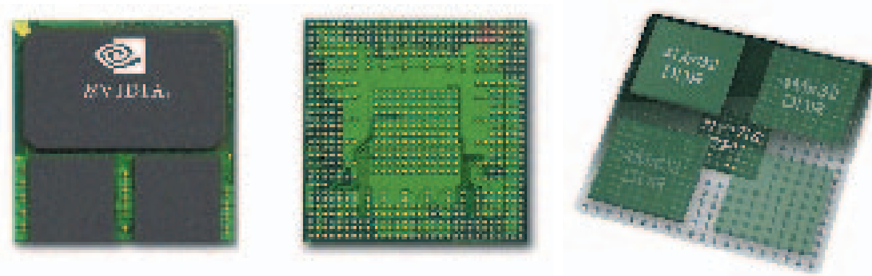
Xilinx
www.xilinx.com

SUPER INFO NUMBER
For more information on the products available from all of the vendors listed in this box, enter no. 324 at www.ednmag.com/info.

Web only



Stacked MCMs minimize the board-space footprint necessary to represent a given function, and both die backgrinding and advanced assembly techniques minimize their height, too (courtesy Sharp).



The GeForce4Go comes in one-, two-, (a) and four-memory (b) configurations. The four-memory connection connects the graphics-accelerator die to the underside of the package substrate (courtesy Nvidia).