THE BALL GRID ARRAY, BGA

Historical Perspective

The concept of soldering a land grid array of pads to a circuit board with solder balls has captivated the industry since the mid 1990’s and has revolutionized the options available for high density and low cost packaging. The desire for an area array solution stems from the basic premise of a smaller footprint for higher pin count packages. This is shown in Figure 10-1. Using a full array for pads on a 50mil grid, a BGA with just 70 leads or more has a smaller footprint than an equivalent 25mil pitch PQFP.

![Figure 10-1. Package Size for Peripheral and Grid Array Packages](source: ICE, “Roadmaps of Packaging Technology”)

This concept is a natural extension of the roadmap of packaging: take a 100mil center PGA package and reduce the grid pitch. First is the interstitial PGA package at about 70mil effective centers. The next finer grid is 50mil centers. However, these small diameter pins are easily bent, making it difficult to produce a circuit board with holes on a 50mil grid that can be used as
through holes for pins, while still allowing clearance between them for routing. 50mil center
PGAs have been fabricated and used in production, but for surface mount only. An example of
such a package is shown in Figure 10-2.

High-end computer systems from IBM, Fujitsu, Hitachi and NEC have used area array bump
packages since the mid 1980’s for single chip packages mounted on high density substrates. An
element is shown in Figure 10-3. Pitches down to 25mils have been used in production for the
NEC SX-X computer. Land grid array single chip packages using an interposer for connection to
a motherboard had a brief appearance. An example of an LGA single chip package is shown in
Figure 10-4.

The latest surge of interest in surface mount array packages has arisen from the pioneering work of
Motorola and Compaq Computer. Motorola, along with Citizen Watch, created the OMPAC, Over
Molded Plastic Carrier, in 1989. This was designed as a low cost, standard style of package, using
a two layer printed wiring board substrate. The chip is attached to the top surface, wirebonded and
overmolded. The bottom side has the solder balls in either a full array or a partially filled array.
An example of the original OMPAC is shown in Figure 10-5. Early designs had 70 to 250 pads and
were used in portable consumer applications because of their small footprint and low cost.

Compaq Computer took the bold step of implementing a 225 pin ASIC in a BGA to offer a reduc-
tion in board area and a potential cost reduction. In 1992, they began full production with BGA
ASICs on PC circuit boards.
Since then, the market opportunities for BGAs have expanded at better than 30% growth rates per year. The predicted market growth is shown in Figure 10-6. By the year 2000, the demand for BGA packages will approach 3 billion units. With an average of 300 pins on each, and an estimated cost of 2¢ a ball, the potential market volume may approach $20B. BGAs will have a profound impact on the near future roadmap of every aspect of packaging technology. These applications will span low pin count single chip packages to complex MCM packages. An example of a complex MCM using a BGA footprint is shown in Figure 10-7.
Figure 10-5. The First PBGA: The OMPAC (Over Molded Plastic Carrier)

Figure 10-6. BGA Market Projections
As with all new packaging technologies, the driving force for rapid acceptance of BGA packages is due to one or more of the following benefits:

- denser packing of functionality
- higher performance
- lower cost

In 1997, just 5 years after the introduction into high volume by Compaq Computer, BGAs are moving into the mainstream throughout the industry. They are in volume production by all the computer companies: HP, DEC, IBM, SGI, Sun, Apple, Compaq, and Motorola. An example of a PowerPC based motherboard is shown in Figure 10-8 with three ASICs in PBGAs and the PowerPC in a CBGA hidden by a heatsink. BGAs are in production with all the ASIC vendors: LSI, Fujitsu, NEC, Toshiba, AT&T, etc. They are also in volume production by all the FPGA vendors.
BGAs are appearing innocuously in many new board designs. An example of a new small size industrial PC board is shown in Figure 10-9. Today, the BGA is firmly established in packaging roadmaps and in a few years will be thought of as just another component.

THE GRID ARRAY FOOTPRINT

All packages having an array of surface mount solder balls or columns has been included in the BGA family. They all share the common feature of an array of solder balls, on area array grid pitches, registered with the JEDEC committee as 1.5mm (59mil), 1.27mm (50mil), or 1.0mm (39mil). This is where the similarities end. The BGAs vary in every other aspect. (In general, packages with a ball grid on a tighter pitch than 1mm are generally termed Chip Scale Packages (CSPs), and are discussed later in this chapter.)

The array of solder balls can be filled or partially filled. As a filled array, every grid point has a solder ball. As an unfilled array, only the outer few rows have balls. An example of each is shown in Figure 10-10. The advantage of the partially filled, or peripheral array BGA is the ease of routing. In a filled array, connections must be made from the innermost balls to the routing channels of the circuit board. In order to access these inner balls, enough layers must be provided in the board to allow the innermost balls to “escape” to the routing channel. This issue of “escapes” is reviewed in Chapter 5.
With one trace able to fit through the array of pads and vias, or one track technology, 2 rows deep of balls can be routed per layer of interconnect. With two traces able to fit between balls, or two track technology, three rows deep of balls can be routed. This is illustrated in Figure 10-11. If there are 625 balls, or 25 rows, typically a 32mm body size, 6 signal layers would be required with one track technology. This simply connects the inner balls to the routing channels.

Figure 10-10. Filled Array and Partially Filled Area PBGA Packages

Figure 10-11. Wireability Requirements for the Substrate PCB
In complex boards that require 6 routing channels for chip to chip routing anyway, there is no penalty for using a filled array. However, a board with fewer layers would have to use either finer lines and two track technology or increase the layer count. Another way of solving this routing problem is to use a partially filled array.

With one track technology, two rows of balls can be routed per layer. Four outer rows of balls is the maximum row depth that can be routed with any conventional board using just two signal layers. The surface layer uses the first two rows, and the second signal layer can use the inner two rows. There is a size penalty for the BGA because the inner balls are missing.

The total number of balls that can be accommodated in a BGA package, if each ball is on a square grid point, is given by:

\[ N = a^2 - b^2 \]

where

- \( N \) = total number of balls
- \( a \) = number of rows per side
- \( b \) = number of rows taken out from the middle

For the removed square of balls to be symmetric, \( a \) and \( b \) must both be odd or both be even numbers.

In this case, the row depth of the remaining balls, \( c \), is given by:

\[ c = \frac{(a - b)}{2} \]

For example, a package might have 19 rows with 9 center rows of balls removed. The total number of balls will be \( 19^2 - 9^2 = 280 \), and the balls would have a row depth of \( (19 - 9)/2 = 5 \). Usually, one ball is removed as a key so the total ball count would be 279. This is illustrated in Figure 10-12.

Leaving only peripheral rows of balls to facilitate routing will increase the size of the BGA package over a filled square. This is shown in Figure 10-13, comparing the size of a PQFP with a BGA with \( n \) rows deep of peripheral pads, where \( n \) varies from 1 to 7. A BGA with 2 peripheral rows, on 50mil centers, is roughly equivalent to a 25mil PQFP in package size.

Even a BGA with just 4 rows (the maximum that any board can route) and 625 balls, will be about half the size and 1/4 the area of an equivalent PQFP, if such a large PQFP could be fabricated. The BGA would be about 55mm on a side, compared to the filled array BGA of 32mm on a side. The partially filled array BGA would have 44 rows with the inner 36 removed.
Figure 10-12. 361 Ball BGA with Center 81 Removed to Result in 280 Ball BGA with 5 Rows of Balls

Figure 10-13. Package Size with Partially Filled Arrays
A 625 ball, filled array BGA is 32mm on a side. It would require 6 routing layers in one track technology, just for escapes. By increasing the size to 55mm, it would require only 4 signal layers. Sometimes the larger package can result in a lower cost system.

Alternatively, filled arrays are used, but the center balls are used for power and ground that do not require escape traces and can still be routed on 2 layers. An example of the routing for a 357 pin BGA with the center 9 rows of balls devoted to ground and the next row devoted to power is shown in Figure 10-14. This design also allocates the outer 4 rows as signal interconnect. One challenge to this design is gaining access to the inner power and ground pads given the necessary connection to power and ground leads around the outer leads.

Figure 10-14. 356 Ball BGA Showing Pin Allocation and Routing on the Substrate

Source: Motorola/ICE, “Roadmaps of Packaging Technology”
BGA FAMILIES

Like all new technologies that quickly proliferate from many different sources in the industry, there is an explosion in acronyms and names for the various specialized versions of BGAs. For example, some of the published types of BGAs are:

- CBGA: ceramic BGA
- CCGA: ceramic column grid array
- TBGA: tape BGA
- PBGA: plastic BGA
- DBGA™: dimpled BGA
- SBGA™: super BGA
- MBGA™: metal BGA
- MicroBGATM: microBGA
- StarBGATM: Star BGA
- FC-BGA: flip chip BGA
- EPBGA: enhanced plastic BGA
- VBGA™: ViperBGA

The great variety of BGA types has evolved, not so much due to performance, but because many of the vendors have a particular specialty in a technology and want to offer a feature that differentiates their package from everyone else’s, or to get around a possible patent barrier imposed by another vendor.

The great variety of BGA package types can be roughly organized onto three different categories based on the material or technology used to fabricate them. The three basic families are:

- Plastic laminate BGA (PBGA)
- Ceramic BGA (CBGA)
- Tape BGA (TBGA)

The cross sections of representatives of each of these are shown in Figure 10-15.

- Each has the same variety of ball counts available
- Each can be used with single chips or multiple chips
- Each can be used with multilayers for enhanced electrical performance
- Each can be used with enhancements for very low thermal resistance and heatsinks attached
- Each can be used with wirebond or flip chip
PLASTIC BGA

Chip Assembly Process

The original PBGA was the OMPAC. The same basic design is still in use. The substrate must be a high temperature substrate to more readily withstand the 150°C thermocompression gold wiring bonding process and the two thermal cycles for ball attach and surface mount reflow. This limits material selection to Polyimides, BT (Bismaleimide Triazine) epoxy-glass and Dry-Clad™, another high glass transition temperature (Tg), low water absorption epoxy based laminate from IBM.
In the lowest cost version, the substrate is a simple double sided circuit board, with the die attach pad and bonding pads on one side and plated through-holes (PTHs) to the bottom side. The bottom side has the fan in from the PTHs to the pads and the solder balls.

The top and bottom surfaces of the package are nickel and gold plated. Gold on the top surface allows gold wirebonding, while gold on the bottom surface assures good wetting by the solder balls. A layer of solder mask is applied to the bottom side of the package to help define the solder pad. There are two generic configurations for the pad, solder mask defined (SMD) and non solder mask defined (NSMD) pads, as illustrated in Figure 10-16.

The general preference is for SMD pads. This allows for less solder to adhere to the pad, keeping the solder ball stand-off height higher. In addition, it has been found that the weakest link in a solder ball joint to the motherboard, is the adhesion of the copper pad to the substrate. The SMD pad configuration has larger copper pads and uses the overlap of the solder mask over the edge of the pad to help hold the pad in place and increase the pull strength of the solder joint. The cross section of these two configuration is shown in Figure 10-17.

The chip assembly and package manufacturing process is shown in Figure 10-18. Die attach adhesive, usually silver filled epoxy, is applied to the laminate, usually in strips. The bare die are die attached to the substrate in strip form. The chips are then wirebonded with gold thermocompression bonding, heating the substrate up to about 150°C. After bonding the molding compound is applied to the top surface.
An alternative protection scheme is to glob top the die. This approach is often used in the prototype phase because it does not require any special tooling and can be performed quickly. An example of a molded and glob top PBGA is shown in Figure 10-19.

The last step is the application of the balls to the bottom surface. This is usually accomplished by applying flux to the bottom surface of the packages and using a vacuum fixture, picking and placing the entire array of preformed solder balls for the entire strip. Usually an optical detector inspects the fixture to verify that there are no missing solder balls. The solder balls are placed on the package, which then enters a conventional reflow oven. A cross section diagram of a typical assembled PBGA is shown in Figure 10-20.

Figure 10-17. Comparison of PBGA Solder Joints

1. Fabricate package substrate in panel or strip form.
2. Ni/Au plate top and bottom pads.
3. Die attach with silver epoxy.
4. Thermocompression wirebond chip.
5. Over mold encapsulate chip.
6. Apply flux or solder paste to package.
7. Place solder balls near fixture for packaging.
8. Reflow solder balls to package.
10. Singulate packages from panel or strip.

Figure 10-18. PBGA Assembly Process
Almost universally, the solder balls used on PBGA are 63% Sn/37% Pb eutectic, melting at approximately 180°C. With eutectic solder balls, the assembly is very robust. There is a growing trend, pioneered by Motorola, to use solder balls of 62% Sn/36% Pb with 2% Ag solder. This solder has a eutectic point at 179°C and offers slightly greater fatigue life.

The packages are then cleaned and singulated. The final step is test and optional burn in, before shipment to the end user or back to the chip fabricator.
Thermal Performance

Thermal performance, as pointed out in Chapter 6, is measured by the junction to ambient thermal resistance, $R_{ja}$, in units of °C/watt. The typical conservative, maximum temperature drop allowed between the junction and ambient is approximately $\Delta T = 50$°C. This corresponds to a junction temperature of 90°C and an ambient temperature of 40°C. Of course, this may vary for specific systems, but is a good general approximation.

The thermal resistance of a package will limit the maximum power dissipation, $P_{\text{max}}$, a chip can have and still meet the 90°C junction temperature. The maximum power handling capability of a package is given by:

$$P_{\text{max}} = \frac{\Delta T}{R_{ja}}$$

This is shown in Figure 10-21. For example, if the thermal resistance is 10°C/watt, the maximum power the package can handle is 5 watts.

![Figure 10-21. Maximum Power Handling Capability for a Package Based on $\Delta T = 50$°C and the Junction to Ambient Thermal Resistance of the Package](source: ICE, "Roadmaps of Packaging Technology"

The thermal performance of the OMPAC style PBGA is strongly dominated by the circuit board substrate it is mounted to. In this respect, thermal performance is an extrinsic property and depends on the board design.
For example, a 225 ball package with a 27mm body size, has a junction to ambient thermal resistance in still air of about 26°C/watt when mounted to a 2 layer substrate. When mounted to a four layer substrate with power and ground planes to carry heat away, the thermal resistance drops to 18°C/watt. When mounted to a 2 layer motherboard, the BGA can support only a 2 watt chip. When mounted to a 4 layer board, it can support a 2.7 watt chip. When the air flow is 4 meters/sec, the thermal resistance when attached to a 4 layer motherboard is about 11°C/watt. This substrate can support a 4.5 watt chip with no heatsink.

When thermal performance is critical, there are five improvements in the package that can be implemented:

- Add balls under the chip just for thermal transfer to the substrate. This can decrease thermal resistance by 25%.

- Add thermal vias to the thermal balls under the die in the package to transport heat more easily to the thermal balls. This can decrease the thermal resistance by another 15%. Figure 10-22 is an example of a cavity-up 479 ball package with thermal vias and balls added to the center for thermal management.

- Add internal metal planes to the package to act as heat spreaders. This can reduce the thermal resistance by another 35%. An example of a PBGA with these three features and its thermal performance is shown in Figure 10-23. For this 35mm PBGA, the thermal resistance in still air was 26°C/watt with no internal metal planes. With 2 internal planes, each of 2oz copper added, causes thermal resistance to drop to 17°C/watt—a 35% decrease in thermal resistance. In all cases, the test was designed with no thermal transfer to the motherboard.
Add a internal metal slug to the package mounting the chip in direct contact with the heat spreader and mounting the package cavity-down. An example comparing a 27mm PBGA and a SBGA™ from Amkor, having a metal base, both mounted on a single layer test board, is shown in Figure 10-24. The thermal resistance, in still air, of the PBGA is 35°C/watt. The thermal resistance of the SBGA, with copper slug, is 22°C/watt. An example of a cavity-down 479 ball BGA package from IBM is shown in Figure 10-25. The metal slug exposed to the top surface is available for heatsink attach.

Add a heatsink to the top of the BGA, either in cavity-up or cavity-down configurations. For cavity-up, shown in Figure 10-26, the thermal resistance for this 27mm package in 3m/sec moving air is 16°C/watt. When a heatsink is added, the thermal resistance drops to 7°C/watt. This PBGA OMPAC style package can easily handle a chip dissipating 7 watts. Using a cavity-down configuration with an integral copper slug and heatsink attached, the thermal resistance can be as low as 4°C/watt for a 31mm package. This would allow a chip with 12 watts. This performance is shown in Figure 10-27.

The UltraSparc processor, operating at over 300MHz, is packaged in a 540 pin PBGA cavity-down package. It dissipates over 30 watts. An example is shown in Figure 10-28. The chip is directly attached to a copper heat spreader, with screws in the back of the heat spreader to clamp a large heatsink. The processor is assembled on a small daughter card module with its L2 cache and bus interface chips.
Grid Array Packaging: BGA and CSP

Figure 10-24. Thermal Performance of Super BGA Compared with Regular PBGA, Using Single Layer Test Board, in Still Air

Figure 10-25. Cavity-Down µLaminate™ Plastic Ball Grid Array (PBGA) with Copper Slug Available for Heatsink Attach
Figure 10-26. Thermal Resistance Versus Air Flow for Plastic BGA with Fin

Figure 10-27. 31mm Cavity-Down BGA with Heatsink
At these powers, the thermal management of the heat flow from the heatsink to the ambient, and managing the air flow in the system enclosure is important. To facilitate high air flow and compact system size, the daughter card is assembled vertically to the motherboard. This is shown in Figure 10-29.

Figure 10-28. 30 Watt UltraSPARC Processor Package in 540 Pin Cavity-Down PBGA on 1.27mm Centers

At these powers, the thermal management of the heat flow from the heatsink to the ambient, and managing the air flow in the system enclosure is important. To facilitate high air flow and compact system size, the daughter card is assembled vertically to the motherboard. This is shown in Figure 10-29.

Figure 10-29. UltraSPARC Motherboard with Processor Card Mounted Vertically to Facilitate Cooling
**Electrical Performance**

Because of the size reduction between the PBGA and the equivalent PQFP, the lead inductance of the PBGA is about half that of the PQFP. In addition, by using vias close to the bonding shelf that extend directly down to the solder ball on the other side, the lead lengths of critical paths, such as power and ground, can be reduced even more. The lead inductance in a 100 pin PQFP can be on the order of 7nH. The lead inductance in a PBGA can be on the order of 2.6nH. This is shown in Figure 10-30.

![Figure 10-30. Lead Inductance of BGA and PQFP](source: Hitachi/ICE, "Roadmaps of Packaging Technology")

In electrically enhanced packages (E-PBGAs), more than two layers are used and power and ground planes can be included. The substrate is more expensive, but the performance is much better.

Substrates with as many as 8 layers have been used for high performance chips, with multiple ground and power planes and 4 signal layers. The power and ground inductance per lead can be as low as 4nH, 2nH for the wirebond and 2nH for the plane and ball. The signal leads can be a controlled impedance of 50 Ohms, continuous to the bond wire pad. This allows extremely good electrical performance. To provide for the complex routing, buried vias and build up multilayer (BUM) technology, described in Chapter 11, can be used. Typically, devices that require the highest electrical performance also require the higher thermal performance, often using a metal slug and cavity-down mounting. An example of the cross section of a multilayer PBGA is shown in Figure 10-31.

EPBGAs are used for the highest clock frequency chips. For example, the Sun UltraSparc processor chip is mounted in a 520 pin 8 layer PBGA with wirebonded leads. It operates at 300MHz.
When multilayer substrates are used in the PBGA package, multiple bonding shelves can be used to wirebond the chip pads. To decrease costs, pads can be located on the top surface only. One commonly used design utilizes power and ground rings around the bonding area of the chip inside of the signal pads. This allows a higher density of power and ground wirebonds and keeps them short. An example of the bonding footprint of an EPBGA package is shown in Figure 10-32, showing the power and ground bond rings.
Package Costs

The total cost of ownership for a package is composed of:

- the part cost, including chip assembly
- the board assembly costs
- the rework cost
- the field failure cost
- the failure analysis cost

The first three are the most straightforward to analyze. It is sometimes misleading to estimate BGA packaging cost using only the cost of the BGA piece part. A study completed by Hitachi, comparing a 208 0.4mm PQFP and a 225 BGA with 1.5mm centers, shown in Figure 10-33, points out that the BGA cost of ownership is slightly lower than the PQFP in this package size range. The increase in piece price of the BGA is offset by the lower rework costs. This is even more pronounced in a 300 pin PQFP. The reduced rework cost of the BGA positions the BGA with a 25% reduction in cost over the PQFP.

The current piece price of BGAs is higher per pin than a 0.5mm PQFP. The price for a 225 PBGA is approximately 2.5¢/pin, compared with 1.5¢/pin for a PQFP. Both prices include chip assembly costs. In OMPAC style PBGAs, the substrate accounts for about 1/3 the total price. For enhanced packages with multilayer substrates, the added price of the substrate materials and yield impact can increase the package price to 4-5¢/pin. A 520 pin BGA at 5¢/pin costs $26 each to purchase in high volume.

Source: Hitachi/ICE, "Roadmaps of Packaging Technology"

Figure 10-33. Total Assembly Cost
A challenge for the vendor base is to create a high performance BGA that can be manufactured at low cost. Two approaches appear promising: the SuperBGA™ from Amkor and the ViperBGA™ from Prolinx Labs.

A cross section of Amkor’s SuperBGA™, or SBGA, is shown in Figure 10-34. The back side of the package is a copper heat slug approximately 10mils thick. On top of this is attached a window frame of a thick laminate substrate. Traces fan out on the top from the bonding pads to the solder balls, defined by solder mask. The chip is attached directly to the copper base and is roughly level with the laminate bonding surface. After the die is wirebonded, it is glob topped. The surface of the glob top is still well below the surface of the solder balls, so the SBGA can be mounted cavity-down.

With no vias, the substrate cost can be minimized. Lead inductance is kept low by the proximity of the internal metal layer. The copper base to which the chip is attached provides very good thermal management.

Prolinx Labs’ ViperBGA™ design is similar but allows for two metal layers, as shown in Figure 10-35. There is one layer of copper in the laminate that acts as a current carrying ground plane. A photoimageable dielectric (PID) is laminated over this and vias are formed. A conductive paste is filled into the holes acting as “microfilled vias”. Copper foil is laminated over the dielectric and holes and traces are lithographically printed. By using the photoimageable dielectric, the cost of drilling is eliminated.

As of 3Q97 manufacturing volumes for both technologies were too low to establish realistic pricing for the SuperBGA™ and ViperBGA™ packages.

**Flip Chip PBGA**

The key feature of BGA packages is their ability to support a high number of package I/O, even over 1,000. For die requiring such a high number of pins, there is a growing trend to use area array solder balls on the chip and to mount them as flip chip. To route the escapes from the 10mil pitch
array of chip solder balls requires a fine via grid on the top surface of the BGA. This is enabled with the use of BUM (build up multilayer) technology, described in Chapter 11. This technology is just coming into production and is based on the use of micro vias and fine line lithography.

In most implementations, a few layers of micro vias and fine lines are fabricated on top of a conventional substrate. An example is shown in Figure 10-36. The chip is flip chip attached by either solder or polymer connection, as described in Chapter 9. For enhanced reliability, the chip is underfilled with a low viscosity epoxy. For lower power applications, the back of the chip can be left exposed to the ambient. For high power applications, a heat spreader can be applied to the back of the die. These two cases are shown in Figure 10-37.
The thermal performance of a flip chip PBGA with and without a heatsink is shown in Figure 10-38. With a 3 meter/sec air speed, the thermal resistance with no heatsink is about 26°C/watt. With a heatsink, this drops to 5°C/watt. Such a package can easily handle a 10 watt chip.

CERAMIC BGA (CBGA)

The ceramic BGA uses the same cofired ceramic technology used for the very popular PGA and MCM modules. This technology is described in Chapter 11. The package is composed of multilayers of molybdenum metallization and alumina dielectric. The vias and traces can be as close as 8mil. This allows pads on the top surface of the package to match the fine pitch needed for flip chip area attach. This was the first application introduced by IBM. An example of the cross section of a FC-CBGA attached to an FR4 circuit board is shown in Figure 10-39.
However, compatibility with the rest of the industry that is currently wirebond-dominated encouraged IBM and other cofired ceramic suppliers to design ceramic BGAs with cavities for wirebonding as well. Some examples of FC and wirebond CBGAs are shown in Figure 10-40.

Figure 10-38. Thermal Resistance of Heatsink Configuration for Flip Chip BGA

Figure 10-39. Exposed Cross Sectional View of a FC-CBGA Mounted to a Substrate
The assembly sequence for CBGAs, illustrated in Figure 10-41, is similar to PBGAs. The multilayer cofired ceramic substrate is fabricated as described in Chapter 11. The chip is die attached, typically with silver filled epoxy for non-hermetic and silver glass for hermetic packages. It is then wirebonded and either hermetically sealed or glob top sealed. Eutectic solder paste is screened on the bottom of the package and 90/10 solder balls placed with a fixture. The solder balls are reflowed to melt the eutectic solder. The chip is then tested and is ready for either burn-in or shipment to the end customer or assembly to the motherboard.

Figure 10-40. Ceramic Ball Grid Array (CBGA) for Wirebonding and Flip Chip

The assembly sequence for CBGAs, illustrated in Figure 10-41, is similar to PBGAs. The multilayer cofired ceramic substrate is fabricated as described in Chapter 11. The chip is die attached, typically with silver filled epoxy for non-hermetic and silver glass for hermetic packages. It is then wirebonded and either hermetically sealed or glob top sealed. Eutectic solder paste is screened on the bottom of the package and 90/10 solder balls placed with a fixture. The solder balls are reflowed to melt the eutectic solder. The chip is then tested and is ready for either burn-in or shipment to the end customer or assembly to the motherboard.

1. Reflow Bumped Die to Ceramic at 370°C – Custom Reflow
2. Probe Test & Epoxy Underfill – Vacuum Dispense
3. Apply Thermal Grease to Die – Screen Dispense
6. Apply Perimeter Silicone Seal – Screen Dispense
7. Apply Lid & Seal – Batch Oven
8. Screen 63/37 Paste – Stencil
9. Apply 90/10 Balls* – Modified Pick & Place
10. Reflow Eutectic Solder – Nitrogen Reflow Furnace

* Note: Balls or Columns. Columns May be Directly Cast onto Package
Source: IBM/ICE, “Roadmaps of Packaging Technology” 22269

Figure 10-41. Ceramic Ball Grid Array Assembly Process Based on the IBM Process
The chief difference between equivalent performance CBGA and PBGA packages is the price. CBGA packages are priced at 5¢-10¢ per pin while the PBGAs are priced at 3¢-7¢ per pin. However, until PBGAs with BUM layers are in production, CBGA may be the only manufacturable approach to achieving the via density needed to route high-pin-count flip chip devices.

**SOLDER JOINT TO CIRCUIT BOARD RELIABILITY**

In addition to the substrate material, the other key difference between the CBGA and the PBGA is the solder ball composition. The ceramic substrate has a close thermal coefficient of expansion (TCE) match with the silicon chip. TCE of ceramic materials is about 6.6ppm/°C, compared with silicon at 2.6ppm/°C. This makes it an ideal substrate for flip chip attach. Adding an underfill of epoxy assures even longer fatigue lifetime.

However, the good TCE match between the ceramic and silicon means there is a poor match to the circuit board substrate’s typically 18-22ppm/°C TCE. The fatigue life of the solder joint stressed from thermal cycling is a major concern with BGA packages.

Solder joints between the package and the substrate fail because of the repeated stress placed on the joints when the two elements expand and contract different amounts as the temperature changes—either through power cycling or isothermally. This is diagrammed in Figure 10-42.

![Figure 10-42. Stresses on Solder Joints Due to Miss-Match in the Expansion Coefficient, During Thermal Cycling](image_url)

Source: ICE, "Roadmaps of Packaging Technology"
The amount of stress the solder joint sees depends on the strain—the displacement between the top and bottom of the solder joint. This strain, $\Delta x$, is related to the distance to the neutral point (DNP), the temperature change, $\Delta T$, and the difference in the TCE of the two materials. The neutral point is the position that does not change through thermal cycling. For a symmetric package, the neutral point is the center of the package. The DNP is the distance to the solder joint. The strain is given by:

$$\Delta x = DNP \cdot \Delta T \cdot \left( TCE_1 - TCE_2 \right)$$

How much stress is created in the solder joints will depend on the compliance of the joint, which is related to material properties, the ratio of the length of the joint, $L$, and the diameter of the joint, $D$. The larger the ratio of $L/D$, the more compliant the joint and the more strain it can tolerate without creating as large a stress.

These two effects, the creation of strain and how it influences the stress on solder joints, governs the solder ball connection design for all BGAs.

**Weibull Distributions and Solder Joint Reliability**

Solder joint reliability is measured by assembling a number of packages on a test board with the solder joints connected in a daisy chain. As the board is thermally cycled, typically from $0^\circ\text{C}$ to $100^\circ\text{C}$, the solder joint resistances are monitored. $-40^\circ\text{C}$ to $125^\circ\text{C}$ is normally used to cover the extreme conditions found in automotive applications.

When one joint shows an open for more than 1 microsecond, it is considered a failure. This usually means a crack has propagated all the way through the solder joint. An example of the cross section of a virgin and failed solder joint is shown in Figure 10-43. A failure analysis is always performed to confirm it was a solder joint failure. In this way, the statistics of how many thermal cycles it takes for each joint to fail can be built up.

Most failures follow a Weibull Distribution. This is of the form:

$$R(N) = \exp\left(\frac{-N}{\alpha}\right)^\beta$$

where

- $R(N)$ is the fraction of devices that have survived after $N$ cycles
- $\alpha$ is the number of cycles for 63.2% of the devices to fail
- $\beta$ is related to the distribution of failures
- $N$ is the number of cycles
Typically 1-R(N) is plotted, the fraction of parts that have failed. Figure 10-44 shows the Weibull distribution for \( \alpha \) of 1,000 and various values of \( \beta \). It is clear that as \( \beta \) increases, the distribution of failures tightens up. A high \( \beta \) means all the parts behave the same and fail at the same time, corresponding with a very well controlled process. Sometimes, the \( N_{50\%} \) is reported, which is the number of cycles for 50% of the parts to fail. When \( \beta \) is large, \( N_{50\%} \) and \( \alpha \) are very close in value.

Figure 10-45 is a plot of the solder joint reliability data for a 225 ball PBGA. \( \alpha \) is 7,958 and \( \beta \) is 13. The distribution is tightly grouped. Figure 10-46 summarizes the failure data for solder joints in a variety of parts.

In particular, note that the \( N_{50\%} \) for a 68 pin PLCC and a 72 ball PBGA, are 6,561 and 2,171, respectively. It is important to keep in mind that the \( \beta \)s are very different—3.3 for the PLCC compared with 9.1 for the PBGA. This means that the distribution of failures for the PLCC is much broader. Failures begin much earlier for the PLCC than the PBGA. In fact, the onset of failures, the \( N_{0.1\%} \) point for the PLCC is 904 cycles, but the \( N_{0.1\%} \) for the PBGA is 1,058 cycles. In terms of product lifetime, the PBGA is at least comparable to the PLCC.

To estimate the in-use lifetime of the joints from the thermal cycling data, the acceleration factor, \( AF \), must be known. This is defined as:

\[
AF = \frac{N_{xx, \text{field}}}{N_{xx, \text{ATC}}}
\]
Figure 10-44. Weibull Distribution for $\alpha=1,000$ and Various Values of $\beta$

Figure 10-45. Log Normal Failure Distribution for 225 PBGA at 0 to 100°C, 2 Cycles Per Hour
In practice, the AF is very difficult to estimate. It can be extrapolated based on the difference in thermal cycle excursion and the scaling of the non-linear behavior of the solder joints, described by the Coffin-Manson Equation. Unless the AF is obtained empirically, or from non-linear finite element analysis using well characterized material properties, it is only an estimate. In general, the AF varies from 2 to 10.

Using a value of 5 for the AF, an \( N_{0.1\%} \) corresponds to 1,000ppm failure rate in about 5,000 cycles for the 225 PBGA. If a thermal cycle occurs once a day in field use, the lifetime for 1,000ppm failures is about 15 years. Values of \( N_{0.1\%} \) of as low as 200 cycles at 0 to 100°C are considered the lower limit of acceptable performance.

### Solder Joint Reliability, PBGAs and CBGAs

The worst case PBGA values are about 1,000 cycles, for a 165 ball PBGA, using the extreme conditions of -25°C to 100°C. A large factor influencing the reliability of cavity-up, full array PBGA joints is the location of the solder ball. When it is located under the periphery of the die, it sees more differential strain and more stress. These balls tend to fail more quickly.

Figure 10-47 shows the Weibull Distribution for balls located on the outer row of the package, the middle of the package and under the die perimeter. The \( \alpha \) of the outer row is 11,468, while the \( \alpha \) for the balls under the edge of the die is 7,958. This is another motivation to remove the balls under the edge of the die and use a partially filled array of solder balls.

---

<table>
<thead>
<tr>
<th>Company</th>
<th>Device</th>
<th>Die Size in MILs</th>
<th>Cycling Temperature (°C)</th>
<th>Number of Cycles Per Hour</th>
<th>( N_{0.1%} )</th>
<th>( N_{1%} )</th>
<th>( N_{50%} )</th>
<th>( \alpha (N_{63%}) )</th>
<th>( \beta )</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorola</td>
<td>72 PBGA</td>
<td>270 x 270</td>
<td>-40 to 125</td>
<td>1</td>
<td>1,058</td>
<td>1,363</td>
<td>2,171</td>
<td>2,260</td>
<td>9.1</td>
<td>Internal</td>
</tr>
<tr>
<td>Motorola</td>
<td>72 PBGA</td>
<td>270 x 270</td>
<td>0 to 100</td>
<td>3</td>
<td>3,013</td>
<td>4,034</td>
<td>6,895</td>
<td>7,222</td>
<td>7.9</td>
<td>Internal</td>
</tr>
<tr>
<td>Motorola</td>
<td>119 PBGA</td>
<td>280 x 270</td>
<td>0 to 100</td>
<td>2</td>
<td>4,459</td>
<td>5,848</td>
<td>9,683</td>
<td>10,113</td>
<td>8.4</td>
<td>Internal</td>
</tr>
<tr>
<td>Motorola</td>
<td>225 PBGA</td>
<td>400 x 400</td>
<td>0 to 100</td>
<td>2</td>
<td>4,685</td>
<td>5,592</td>
<td>7,737</td>
<td>7,958</td>
<td>13.0</td>
<td>Internal</td>
</tr>
<tr>
<td>Motorola</td>
<td>361 PBGA</td>
<td>450 x 450</td>
<td>0 to 100</td>
<td>2</td>
<td>6,319</td>
<td>7,804</td>
<td>11,495</td>
<td>11,886</td>
<td>10.9</td>
<td>Internal</td>
</tr>
<tr>
<td>Motorola</td>
<td>68 PLCC</td>
<td>Cu Leadframe</td>
<td>-40 to 125</td>
<td>1</td>
<td>904</td>
<td>1,818</td>
<td>6,561</td>
<td>7,332</td>
<td>3.3</td>
<td>Internal</td>
</tr>
<tr>
<td>Motorola</td>
<td>208 PQFP</td>
<td>Cu Leadframe</td>
<td>-40 to 125</td>
<td>1</td>
<td>639</td>
<td>1,553</td>
<td>7,912</td>
<td>9,111</td>
<td>2.6</td>
<td>Internal</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>169 PBGA</td>
<td>364 x 359</td>
<td>0 to 100</td>
<td>2</td>
<td>2,103</td>
<td>2,741</td>
<td>4,459</td>
<td>4,651</td>
<td>8.7</td>
<td>Semi/HDP-10/93</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>225 PBGA</td>
<td>389 x 398</td>
<td>0 to 100</td>
<td>2</td>
<td>1,412</td>
<td>1,993</td>
<td>3,749</td>
<td>3,960</td>
<td>6.7</td>
<td>Semi/HDP-10/93</td>
</tr>
<tr>
<td>Compaq</td>
<td>72 PBGA</td>
<td>270 x 270</td>
<td>-25 to 100</td>
<td>2</td>
<td>1,734</td>
<td>2,194</td>
<td>3,379</td>
<td>3,508</td>
<td>9.8</td>
<td>1993 IEPS</td>
</tr>
<tr>
<td>Compaq</td>
<td>165 PBGA</td>
<td>437 x 437</td>
<td>-25 to 100</td>
<td>2</td>
<td>1,010</td>
<td>1,309</td>
<td>2,106</td>
<td>2,195</td>
<td>8.9</td>
<td>1993 IEPS</td>
</tr>
<tr>
<td>Compaq</td>
<td>225 PBGA</td>
<td>389 x 398</td>
<td>-25 to 100</td>
<td>2</td>
<td>1,252</td>
<td>1,664</td>
<td>2,807</td>
<td>2,937</td>
<td>8.1</td>
<td>1993 IEPS</td>
</tr>
</tbody>
</table>

Source: Motorola/ICE. "Roadmaps of Packaging Technology" 22273

Figure 10-46. PBGA Accelerated Thermal Cycling Data from Motorola and Others (With Comparisons to PQFP/PLCC)
For the same solder ball geometry, the stresses on a solder joint would be worse for a ceramic BGA than a PBGA because of the larger TCE mis-match between the ceramic and the typically FR4 substrate. To compensate for this mis-match, and to increase the solder joint lifetime, a few techniques have been applied to CBGAs.

The solder ball height is increased by using a non-collapsing solder ball. In CBGAs, the solder ball is exclusively 90% Pb/10% Sn. The melting point of this solder is 300°C, well above the 180°C eutectic temperature. The high temperature solder ball is attached to the ceramic package and to the substrate using a layer of eutectic solder, but the ball acts an inert standoff.

Motorola has compared the solder joint reliability of a 256 lead CBGA with 62% Pb/36% Sn/2% Ag solder balls and 90% Pb/10% Sn solder balls, using the same initial solder ball diameter of 35mils. When the high temperature solder is used, the final standoff height is 37mils. When the eutectic solder is used, the standoff height collapses to 21mils. The solder ball geometry is shown in Figure 10-48.

The increase in height of only 16mils for the high temperature solder ball increases its N50 almost a factor of 3, from 712 cycles to 1,944 cycles, with a temperature range of 0°C to 100°C. Based on similar experiments, all CBGA packages use high temperature solder balls that do not collapse. This drives reliability closer to that of PBGAs.
Increasing the Reliability of CBGAs

As the size of a package increases, the DNP for solder balls on the perimeter increases, and the lifetime decreases. Even with high temperature solder balls, 32mm is the largest body size that has acceptable solder joint lifetime. The DNP is half the diagonal or 21mm. The $N_{1\%}$ for a $-40^\circ C$ to $125^\circ C$ temperature cycle is on the order of 300. A body size of 32mm corresponds to a 25 x 25 full array, with a total of 625 balls. To build a larger package, IBM has resorted to using columns of high temperature solder, rather than balls, to increase the stand-off even further.

Figure 10-49 is an example of a 361 CBGA and a 625 ceramic column grid array package (CCGA). The columns are 20mils in diameter and vary from 50mils to 87mils in height. This is reminiscent of the surface mount PGA on 50mil centers. The difference is that the solder columns can be molded in place. However, they are just as fragile.

The increase in solder joint lifetime for the CCGA is dramatic. The $N_{1\%}$ for a 625 CBGA was measured by Motorola to be 298 cycles, with $0^\circ C$ to $100^\circ C$ excursions. With a 625 CCGA, the $N_{1\%}$ was increased to 1,073—over a 3x improvement. An example of a solder column after 3,000 thermal cycles is shown in Figure 10-50. The failure is in the high temperature solder, near the top and bottom base, where shear forces are highest.
Figure 10-49. Ceramic Grid Array Packages

Figure 10-50. Crack Propagation of a CCGA Solder Joint, After Thermal Cycling
Based on the Weibull Distribution data, and making some assumptions about the acceleration factors, Motorola estimates the field life of a CCGA to be over 2,000 cycles with a 40° thermal excursion in the field, and a 100ppm failure rate. This is shown in Figure 10-51, comparing the field failure rate expected for a 625 CBGA and a 625 CCGA.

A second approach to increasing the height of the solder column is the use of a “dimpled” solder ball developed by Kyocera. In their cofired ceramic BGA packages, an extra layer of ceramic is applied to the pad side, with holes 0.2mm deep and 0.55mm in diameter. The holes are filled with eutectic solder paste by screen printing and solder balls are placed on top. After reflow, the ball has a dimpled shape. A cross section of a dimpled ball is shown in Figure 10-52.

The solder is an alloy of Sn/Pb/Bi, with a melting point of 183°C. This means it can be assembled in the normal SMT flow. The combination of solder composition and dimple shape extends the fatigue life of the CBGA solder joint by almost a factor of 2. For example, for -40°C to 125°C, a 35mm CBGA has an N50 of about 300 cycles. With dimples, the same size DBGA has an N50% of 500. The reliability of smaller sized CBGAs is also significantly extended. Figure 10-53 compares the relative lifetimes for some of these CBGA packages.

Figure 10-51. End-of-Life Projections for a 32mm Fully Populated Solder Grid Array (DND=21.55mm)
Figure 10-52. Structure of a Dimpled CBGA Solder Ball

Figure 10-53. Board Level Reliability of Dimpled BGA
The third type of BGA is the Tape BGA (TBGA). In its various forms, a single or double layer of fine line flexible circuitry forms the inner row of bonding pads and the array of solder balls. An example of a large pin count TBGA is shown in Figure 10-54.

The advantage of using flexible interconnect is the ability to achieve fine lines. Typical features are 35 micron traces on 75 micron pitches. For wirebond applications, this allows a row of pads on the bonding shelf that can match the chip pads. For flip chip applications, a single layer of interconnect can route 3 rows of pads. Two layer tape can route 6 rows deep.

In the construction of the package, a copper heat spreader acts as the base to which the flex tape is laminated. For wirebonded die, the chip is mounted in a cavity in the heat spreader and wire-bonded to the pads on the top surface of the tape. After assembly, the die is molded or glob topped. In wirebond applications, the TBGA inherently has a cavity-down orientation. A wire-bonded TBGA is shown in Figure 10-55.

For flip chip or TAB attach applications, the chip is attached to the opposite side of the tape as the solder balls. When assembled to the board, the chip is cavity-up, and the back side of the die is available for heatsink attach. A cross section of a TBGA with a die attached using a TAB inner lead is shown Figure 10-56.
In all TBGA configurations, thermal performance can be excellent. The back of the die can be attached to a copper heat spreader, exposed to the ambient or attached to a heatsink. The junction-to-case thermal resistance can often be as low as 0.6°C/watt. With heatsink attached, in 2 meters/sec of air flow, the junction to ambient thermal resistance can be as low as 4°C/watt, capable of handling 12 watt chips easily. Figure 10-57 shows the thermal performance of a TBGA package with exposed die, just heat spreader and with heatsink attached.

When a heat spreader or ground plane is used to carry current, the inductance of the ground path can be less than 3nH. Signal lines can be configured as controlled impedance microstrip, right up to the pad, with exceptionally short wirebonds. Good signal integrity has been obtained with rise times as short as 200psec.

There are two methods of attaching the solder balls. In the original IBM implementation, the solder balls are 90/10 high temperature solder balls. Each ball is individually spot welded to through vias in the flex film. The solder ball only reflows slightly, allowing solder to flow through
the hole and form a rivet. This locks the solder ball in place. After the stiffener is applied to the back surface, the TBGA is assembled to the substrate as any CBGA, by applying solder paste to the substrate and reflowing the eutectic solder to the high temperature ball. The solder ball does not collapse.

ASAT and 3M have independently implemented an approach closer to the PBGA. The package is assembled with the chip wirebonded and encapsulated. The TBGA has pads defined by the solder mask. Flux is applied to the pad surface and eutectic solder balls (62/36/2) are placed by mechanical fixture and reflowed. During assembly to the substrate, the solder balls will collapse, just like a PBGA. This assembly process sequence is outlined in Figure 10-58.

Olin has implemented a slightly different version of a TBGA that does not use a free standing flex or tape layer, but essentially results in the same construction. Figure 10-59 is a cross section view of the metal BGA (MBGA). It is fabricated by MMS, using thin film, MCM technology. In its simplest form, an aluminum wafer is used as the substrate. An 18 micron layer of anodized aluminum is grown, providing a dielectric layer. A single layer of thin film copper, nickel and gold is patterned. The features can be as fine as 20 micron lines on 50 micron centers. Conventional solder mask is applied on the top surface to define the pads for solder balls. Later versions might use polymer layers such as polyimide or BCB as the dielectric. This will facilitate multilayer designs and will use lower dielectric constant materials.

After the wafer of packages is fabricated, each one is singulated and the assembly process is continued. The die are silver filled epoxy die attached, and wirebonded. The thin film assures a pitch of the bonding pads that can match the pitch on the chips. For flip chip operation, a 50 micron routing pitch means five rows deep of solder balls on the chip can be routed.
Figure 10-58. TBGA Assembly Process Flow

Figure 10-59. Cross Section of a Single Metal Layer MBGA™ Package
The thermal and electrical performance of the MBGA is comparable to that of the other TBGA versions. The die can be in direct thermal contact with a very good heat spreader, wirebond lengths can be kept short and traces can offer controlled impedance. For additional cost, vias can be generated in the dielectric and the aluminum base used to carry return current. This decreases the ground inductance to less than 2nH.

A significant development in the thin film processing may position the MBGA to be moving toward lower cost. MMS has recently introduced a large area panel (LAP) process. An example of their 16 inch square panel with an array of MBGAs is shown in Figure 10-60. Instead of using 6 inch round wafers, they can process MBGA packages on panels 16 x 16 inches in size. This is the beginning of the long heralded transition for thin film interconnect substrates to be fabricated in panel form, which offers a cost structure closer to PCBs, rather than in wafer form with a cost structure similar to ICs.

ASSEMBLY TO THE BOARD

An initial fear about BGAs among the board assembly vendors stems from the fact that once assembled, the solder joints cannot be inspected. The tradition in surface mount assembly historically has been to perform an inspection of all the peripheral joints on PQFPs. As the pitch
gets finer, the number of solder bridges increases, and the need for inspection is important. When bridges are found, the peripheral leads of the QFP are easily reworked with a fine tip soldering iron.

This technique cannot be used with BGAs. The lack of inspectability scared many assembly vendors and limited the early growth rate of BGAs. However, the early work of IBM, Motorola and Compaq demonstrated that in fact, the first pass assembly yield of BGAs was actually much higher than for QFPs and with a properly defined process, BGA assembly is more robust than QFP assembly.

The generally accepted solder defect rate for 0.5mm (20mil) QFP leads is 50-100ppm per lead. This leads to a failure rate of approximately 2% of devices for a 208 PQFP. For a board with 10 QFP components, it is expected that 1 in 5 boards will require rework to fix a solder joint problem, typically a short between adjacent fine leads.

The assembly yield for BGA leads is less than 1ppm. For a 256 lead BGA, the defect rate will be approximately 200ppm per package. On a board with 10 BGAs, 1 in 500 boards might require rework.

Of course, the assembly yield is a moving target. As an assembly line gains experience with a process, the yield is expected to increase. The evolution of the assembly yield at an IBM facility is shown in Figure 10-61. In 1994, the defect rate for a 361 lead CBGA was 1ppm per lead. For a 168 lead, 0.3 mm pitch PQFP, it was 30ppm.

![Figure 10-61. QFP Solder Defect Summary](image-url)
There are two reasons why the BGA assembly process is so robust and the defect rate can be so low. First is the large spacing between the leads. It is very difficult to create solder bridging, unless the components are severely warped or an extremely large amount of solder paste has been used. The 25mil gap between pads requires a lot of solder to bridge. Secondly, the surface tension of the solder is so high that as long as the package gets within the vicinity of the pads, it will self align. Misalignment can be as high as half a pad width, 12mils, with no impact on defect rates.

Mechanical engineers often say that every widget should be designed so that it can be assembled by putting all of its parts in a bag, shaking up the bag and having the assembled unit fall out. Assembling BGAs to a board comes close to this model.

Solder Paste and Reflow

For PBGAs, it is recommended that boards be screen printed with solder paste. Generally, the solder paste height is 6mils. This height assures that every ball makes contact with the solder paste. The JEDEC spec for coplanarity of solder balls is 6mils. However, Motorola and Compaq have reported that there is no yield change for PBGAs with coplanarity varying from 4mils to 10mils.

For CBGAs, the solder paste plays a far more critical role. The solder paste applied to the board is the only thing holding the CBGA on. The solder volume must be between 5,000mil$^3$ to 10,000mil$^3$, with 7,000mil$^3$ nominal. If the aperture is 30mils in diameter, this corresponds to about 10mils of solder paste for the nominal thickness.

After solder paste screening, the BGA components are picked and placed using conventional SMT equipment. An example of the placement of a large filled array PBGA is shown in Figure 10-62. The boards, with all the other SMT parts placed, is transported though a reflow oven, using a standard temperature profile. An example of a profile is shown in Figure 10-63.

Though there has been a lot of discussion about the use of more exotic reflow processes, such as vapor phase, standard reflow ovens are typically used with a standard temperature profile. A distinction for BGAs is that the temperature of the center of a BGA package must be measured when setting up the zone temperatures. The extra thermal mass of a large package and the thermal insulation of the center ball from the outside air may require adjustment of the zone temperature to achieve the normal temperature profile.

Because of the tacky nature of the solder paste, even double sided BGA boards can be assembled, providing the BGA packages are lightweight. Figure 10-64 is an example of a double sided DIMM module with 4,119 BGAs on each side. After the parts are placed on one side, the board is flipped and parts are assembled on the bottom side. Then the whole board is reflowed. During liquidus, the surface tension of the solder is strong enough to hold the upside-down BGA packages on its bottom side.
Figure 10-62. Pick and Place of PBGA Component to Circuit Board

Figure 10-63. A Reflow Temperature Profile for BGA Assembly
In the early stages of setting up a process, inspection of the BGA joints can be performed using x-ray imaging. An example of an x-ray image of a 225 ball PBGA is shown in Figure 10-65. This particular part suffered popcorning and the center balls show shorts and possibly an open.

**Voids in Solder balls**

Upon closer inspection of the balls, many voids are seen, as shown in Figure 10-66. Voids in solder balls are very common. Much controversy exists about their value or detriment. They are caused by gases that evolve from the solder flux reacting with surface oxides on the solder that are not able to diffuse through the flux and solder during the reflow process.
The finer the solder microspheres in the solder paste, the more surface area they have and the more gas is evolved from reaction with the oxides on the surfaces, thereby increasing voiding. Also, the higher the melt temperature of the flux, the more viscous it is during reflow and the harder it is for bubbles to escape to the surface of the solder ball, creating more voids.

To minimize the presence of voids, a solder paste with a low melt point flux, or at least low viscosity and large solder microspheres can be used, with a reflow temperature profile that is longer at liquidus to allow the voids to bubble to the outer surface and escape.

Voids can coalesce inside the solder ball and grow. Figure 10-67 is an example of the cross section of a solder ball showing a void. In solder joint reliability studies performed by Motorola and Compaq, no impact on the fatigue life of joints has been observed from the presence of voids. However, the generally used acceptance criteria for voids is less than 1/4 the diameter of a ball.

**Special Handling for BGA Boards: Solder Mask and Moisture Sensitivity**

If through hole components are also on the board and wave soldering is the next step, the BGA components must be protected so they do not see a second, partial reflow. As the bottom of the board goes through the wave, the through vias routing inner rows of balls can wick heat up to the BGA solder balls and cause them to reflow. If the board also flexes slightly as it goes over the
wave, opens can result. One way of preventing the heat transfer to the solder balls is to coat the vias on the bottom of the board with a peelable solder mask, as thermal insulation. An example of the bottom of a board with a screen printed, peelable solder mask is shown in Figure 10-68.

PBGA packages are also sensitive to moisture. Because the die are typically larger than PQFP devices, having more I/O, there is a larger area under the die that can trap moisture. During the reflow process, trapped water can turn to steam and crack the die. A bubble formed under the die can warp the package and distort the solder balls, causing opens and shorts. This effect is called popcornning, as it is the same process that creates popcorn when corn kernels are dropped in hot oil.

To characterize the moisture sensitivity of packages, JEDEC and the IPC have introduced standards. The JEDEC version is shown in Figure 10-69. A package classified as Level 1 means that it can stand being exposed to 30°C and 90% RH, presumably the worst shop floor environment, for an unlimited amount of time and not suffer any problems during repeated reflow conditions.

The typical moisture sensitivity of most small BGAs is Level 3. This corresponds to a maximum exposure time of 1 week on the shop floor, without any danger of popcornning during reflow. If a package is exposed for longer than it is qualified, it must go through a bake, which is typically at 125°C for 12 hours, before it can be elevated to reflow temperatures. While in inventory, boards should be kept in desiccators and the parts should be sealed in dry bags.
Level 3 introduces minimal changes to the process flow. Parts are removed from their bags daily, assembled on boards, and the boards are processed through test and rework within a week. Some parts are rated at only level 5, able to be exposed for only 24 hours. These require special handling.

Of course, CBGAs do not have a sensitivity to moisture, and meet Level 1 conditions.
Rework

After solder joint inspection, BGA replacement is the next delicate topic for board assemblers. Many of the same techniques used for fine pitch SMT can be applied to BGA replacement, making it a routine process with multiple vendors for the replacement equipment. An example of two BGA rework stations is shown in Figure 10-70 and Figure 10-71.

![Low Cost BGA Repair Station BGA-2000 Series Rework System](source: OK Industries/ICE, “Roadmaps of Packaging Technology”)

**Figure 10-70. Low Cost BGA Repair Station BGA-2000 Series Rework System**

A typical BGA replacement process is shown in Figure 10-72. If the BGA part is to be sent to the failure analysis (FA) lab, and it has exceeded its exposure to the environment, the entire board must be baked out. With most rework equipment, the heating from removing and reforming a BGA is localized so there is no danger of adjacent components being reflowed.

A vacuum head is lowered over the BGA package and hot air is blown through a nozzle to bring the part up to liquidus. The nozzle acts as a micro oven. An example of the air flow pattern is shown in Figure 10-73. The hot air heats the top side of the package and is vented up and away from nearby components. Each nozzle is customized for every BGA part, so as to minimize the clearance needed to fit between components.

The temperature profile that can be obtained is very similar to a reflow oven profile, with approximately 75 seconds above liquidus. A typical profile is shown in Figure 10-74. A vacuum wand on the inside of the oven then lifts the BGA with its molten solder balls. The removed BGA is either discarded or reprocessed with new solder balls to go to test and then failure analysis.
To replace a new BGA on the board, the site must be redressed. First, the residual solder is removed, using either a soldering iron and solder wick or a hot air/vacuum gun. One nozzle blows hot air on the site to melt the solder. The other nozzle provides vacuum for the hot air and molten solder. In some designs, these nozzles are coaxial.
For PBGAs, it is not essential that new solder paste be applied to the site. For CBGAs, it is. This can be accomplished in a number of ways. If there is sufficient clearance, a microstencil can be placed over the site and solder paste screened through it with a doctor blade. Alternatively, solder paste drops can be dispensed individually through a nozzle. An example is shown in Figure 10-75. Another method is to place the new BGA ball side up in a fixture that will allow solder paste to be screened on top of the balls.
The BGA to be assembled is then positioned above the site on the board and precision aligned. The most routine way of doing this is with split optics that looks up at the base of the BGA and down at the pads on the board. By superimposing and displaying both images, an operator can move the BGA to align the balls with the pads. An example of what is seen with poor and with good alignment is shown in Figure 10-76.

![Figure 10-75. Solder Paste Dispense on Re-Dressed Site of Printed Circuit Board Prior to Placing New BGA](image1)

![Figure 10-76. Beam-Splitter Vision System](image2)

The new BGA is then lowered in contact with the board and the reflow temperature profile is re-applied. The nozzle is then lifted and the board is tested.
This technique for BGA rework has been successfully applied in all the contract and captive assembly houses that work with BGA devices.

When the assembly defect rate is 1 lead per million, and the BGA has 300 pins, about 0.03% of the packages might have a bad solder joint. However, with high pin count ASICs, there is a much higher percentage of parts that will fail in the system due to chip defects. This is especially true for high pin count ASICs and processors. To allow the chip vendor to move up the learning curve and decrease the tester escapes, these bad dice must be captured, failure analysis performed and corrective action taken.

Once a part is removed from the board, it is preferred that it be re-balled so that the same test equipment used to test the original BGA can be used, and the failed part possibly re-mounted to a board to explore the failure mode.

There are a number of simple processes to re-ball a BGA. Once removed from the board, the solder on the package can be cleaned up with the same process used to clean motherboards. For a large volume of parts that must be re-balled, mechanical fixtures to pick up and place the solder balls can be used, similar to that used in the original fabrication of the BGA. For low volume, a solder ball carrier from Raychem can be used.

Raychem’s process is outlined in Figure 10-77. The cleaned BGA is coated with solder flux (a). A carrier, with solder balls positioned in it with the same pattern as the BGA (b), is placed over the package and clipped in place. The assembly is reflowed either in a reflow oven or just with a hot air gun. By applying water to the carrier, it can be peeled off, leaving a re-balled BGA (c). The flux is cleaned off and it is ready for test, or even re-assembly.

CHIP SCALE PACKAGING (CSP)

Definition

In the late 1980’s, the “buzzword” in the packaging industry was MCM. As we illustrate in Chapter 12, it has taken almost 10 years, but MCMs are now established as a viable mainstream packaging technology and versions are in volume production for applications as diverse as super-computers and portable cameras.

In the early 1990’s the buzzword was BGA. It took less than 5 years for BGAs to move from a popular discussion topic to firm establishment in the industry as a mainstream packaging technology.
The next new technology in the packaging roadmap—the buzzword of the mid 1990’s—is Chip Scale Package (CSP). A CSP is popularly described as a package close to the size of a chip, but without the fragility and interconnect problems associated with a bare chip. The motivations to look for solutions in this form factor are the same ones that drive the evolution of the packaging roadmap: finding solutions that are denser, faster or cheaper.

The distinction between a conventional single chip package and a CSP is arbitrarily taken as when the package area is no more than 1.2x the chip area. This corresponds to a packaging efficiency of 83% or greater for a CSP. There is some ambiguity in the industry of whether the package has to be less than 1.2x the chip area or 1.2x the chip length. This is a difference between a packaging efficiency of 83% versus 69%. The most popular usage is for a packaging efficiency of 83% or higher. In either case, the package footprint is only slightly larger than the chip.

CSP designs fall under two headings, based on their lead form factor: peripheral and area array. The various technologies proposed for CSP implementations are summarized in Figure 10-78.

---

**Figure 10-77. BGA Re-ballig Process**

1. Remove residue solder from BGA by: solder wick, hot air/vacuum gun, hot air knife.
2. Flux BGA.
3. Align solder ball carrier over BGA pads and clip.
4. Send through reflow oven.
5. Wet solder ball carrier and peel off.
6. Clean flux from BGA package.
Peripheral CSP

The condition for a peripheral package to qualify as a CSP is:

\[
\left( \frac{L_{\text{chip}}}{L_{\text{package}}} \right)^2 > 0.8
\]

or

\[
L_{\text{package}} < 1.1 \cdot L_{\text{chip}}
\]

This is a difficult condition to be met, except for a few examples. A pad limited die would have a pad pitch on the order of 4-5mils. A CSP package with peripheral leads would have to have a pitch on the order of 4.4-5.5mils. Most practical peripherally leaded CSP versions would be appropriate to non-pad-limited die, such as some memory and many discrete or SSI (small scale integration) chips. This is a huge market by itself. In 1996 alone, it is estimated that the unit volume of discrete diodes, transistors or low pin count SSI devices is over 200 billion units, and growing. This trend is shown in Figure 10-79.

There are only three viable peripheral CSP technologies being proposed in the industry.

One approach has been around for a while, taking advantage of flip chip TAB. The leads come out with little fan out, matching the pitch on the chip.
Another approach currently in volume production for memory chips is termed lead-on-chip (LOC). An example of a LOC design is shown in Figure 10-80. All high density memory chips have the I/O pads located along a center strip of the die. A leadframe is mounted over the die with the tips of the leads very close to the pads on the chip. The chip pads are wirebonded to the tips of the leads, and the entire assembly is molded. The outer form factor looks like a SOJ package.

The third approach, available from Micro SMT, is similar to the old AT&T approach of “beam leads”. Attachment of the peripheral leads is performed in wafer form, utilizing the scribe lines between chips. Beam leads are metalized from the die pads into the scribe lines. The silicon is etched from beneath them, turning them into air bridges. The top of the wafer
is encapsulated and backside lapped, exposing the trenches and creating flexible beams. The wafer is sawn and the chip attached to the substrate by the beams. This process and an assembled unit are shown in Figures 10-81 and 10-82.

This process is being proposed for low pin count devices, less than 100 pins, used with discretes and SSI chips. A possible advantage is the lower cost, due to wafer processing of the package.
Area Array CSP

An area array format with tighter pitch and small body size is a natural extension of the grid array format of PGA and BGA. In the BGA regime, there are three grid pitches that JEDEC has standardized: 1.5mm, 1.27mm and 1.0mm. As part of the roadmap for finer pitch grid array packages, JEDEC and the EIAJ have proposed additional pitches as standards:

<table>
<thead>
<tr>
<th>JEDEC</th>
<th>EIAJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 mm</td>
<td>1.0 mm</td>
</tr>
<tr>
<td>0.8 mm</td>
<td></td>
</tr>
<tr>
<td>0.75 mm</td>
<td>0.65 mm</td>
</tr>
<tr>
<td>0.5 mm</td>
<td>0.5 mm</td>
</tr>
<tr>
<td></td>
<td>0.4 mm</td>
</tr>
</tbody>
</table>

The range of grid pitches at 0.4mm (10mil) and below are considered to be bumped die or flip chip pitches.

With the introduction of the term CSP, there is a continuum in grid pitches from the 100mil pitch of the PGA down to 2mil grid pitch on chip. This is illustrated in Figure 10-83. In each of the package ranges, there is a variety of technologies to achieve essentially the same form factor, each with a different cost structure and associated risk.

The migration from the coarse pitch of 2.5mm to the finest pitches below 0.4mm is driven by the same basic forces that affect the packaging roadmap, the desire for denser, faster, cheaper.
All the different approaches for area array CSP can be categorized into three families based on the interface between the chip and the substrate: direct attach, flexible interposer and rigid interposer. They all can provide a grid array form factor meeting the JEDEC or EIAJ spec that is independent of the pads on the chip. As the die shrinks or different vendors’ chips have different pinouts, the various approaches can provide a fixed format of pads. For some technologies, the pads must be contained within the dimensions of the chip. In others, a frame is used that would allow the balls to extend outside the chip size.

The direct attach method is similar to wafer bumping, often with a redistribution layer to provide a change in the pitch from the pads on the die to an area array. Some of these processes are performed typically on the whole wafer which has a cost structure comparable to wafer bumping. This is the technique used by GE, Sandia and Mitsubishi, in addition to all those companies using just flip chip.

The process GE uses is outlined in Figure 10-84. It takes advantage of their “chips first” overlay technology of building an interconnect layer on top of either a wafer or collection of individual die mounted to a common substrate. Though it has been used mostly for MCM fabrication, as described in Chapter 12, it can also be used to create CSPs for individual die. An example of applying the process to different die, yet retaining a common footprint in shown in Figure 10-85. A CSP device fabricated with the GE overlay process is shown in Figure 10-86.
The CSP design of Sandia, the Mini-BGA, is shown in Figure 10-87. It uses a redistribution layer that can be applied to the wafer. A large solder ball is applied to the top surface.

Mitsubishi’s process uses a redistribution layer on the chip to route pads to standardized positions. Under-bump metalization is applied to the pads to build up a layer of 95% Pb/5% Sn solder, to a depth of 40 microns.

Next, a 100 micron high inner bump of Ni/Au coated copper, previously deposited on a film carrier is transferred to the solder pads of the chips. The solder is reflowed at 350°C to solder the copper bumps in place on the chip pads. The chips are then transfer molded and encapsulated using epoxy resin. On the top surface of the copper bumps, eutectic solder balls are placed by mechanical fixturing, and reflowed. The chip is then ready to be assembled to the substrate. The CSP ready for assembly is shown in Figure 10-88.
Figure 10-85. HDI Chip-on-Flex Chip Scale Package with Common Pad Array Configurations for Different Chip Suppliers

Figure 10-86. HDI Flexible Circuit CSP
The rigid interposer method is a direct extension of the BGA package. The distinction is that it allows an extension of ball pitch down to the CSP regime. The interposer method is used by Amkor, Sharp, Sony, Motorola, IBM, Toshiba and Matsushita, for example.

Amkor’s ChipArray BGA and Sharp’s CSP use a double sided substrate very similar to the OMPAC style, with the chip wirebonded on the top surface. Some examples are shown in Figure 10-89. Sony is shipping 15,000 DCR-PC7 Camcorders per month using CSP components having up to 166 balls on 1.0mm pitch.

Figure 10-87. Sandia’s Mini-BGA

Figure 10-88. Mitsubishi Chip Scale Package
Motorola’s version, the SLICC (slightly larger than IC carrier), is used with flip chip die attach. A cross section is shown in Figure 10-90. A laminate acts as the interposer to fan out the chip’s bump pitch, on the order of 0.4mm pitch, to a pitch closer to the substrates’ of 0.75mm or 1.0mm pitch. The chip is underfilled to increase the fatigue life of the solder bumps.

IBM and Toshiba use a cofired ceramic substrate as the interposer for their CSP. In the Toshiba design, shown in Figure 10-91, a single layer of metalization provides the redistribution layer for the chip array to the package array. The chip is also underfilled. The CSP pad pitch is 0.8mm and 1.0mm.
Tessera CSPs

The most visible vendor for CSP is Tessera. The company’s business plan is to license their technology to assembly, package and chip vendors, while providing the technical support and guidance to bring up their customers’ production lines. Amkor, Shinko, Intel, AMD, TI, and Hitachi are among the licensees who are ramping up production volume.

The name micro-BGA™ is trademarked by Tessera. Like many aspects in the packaging industry, there is some ambiguity about the term and it is sometimes used to describe all CSP with a ball pitch less than the 1.0mm pitch of the finest BGA. ICE suggests its use in reference to Tessera CSPs only.

The micro-BGA design uses a flexible interposer between the chip surface and the solder ball surface. A cross section of Tessera’s design is shown in Figure 10-92. Figure 10-93 shows the assembly process flow for the micro-BGA. A flex film with an elastomer base and 50 micron lines and spaces is laminated to the surface of a chip with conventional peripheral pads. Gold plated tabs extend from the traces over the edge of the film, just above the chip pads. A single point
A thermocompression bonder is used to bond the tab lead to the chip pad. The bonded chip is encapsulated and the solder balls applied and reflowed to the top of the tape. After singulation, the micro-BGAs are ready for test or shipment.

The flexibility of the gold lead takes up the TCE mismatch between the chip and the substrate, as shown in Figure 10-94.

The top of a die with a Tessera overlay is shown in Figure 10-95. The back of the die can be encapsulated for protection or left bare for heatsink attachment. In this basic design, the balls fan in from the periphery of the chip pads. Many variations are possible from this basic concept. An extension can be added to the overlay to allow the balls to also fan out. An example of this is shown in Figure 10-96, from Shinko. Two metal layers can be used in the tape to provide a ground plane for extremely low inductance.

Tessera is probably the closest to volume manufacturing with their approach. By using a film overlay, they can take advantage of the low cost of reel to reel processing. Figure 10-97 is an example of a strip from a 35mm reel of a 46 pin CSP. When 70mm wide tape is used, as shown in Figure 10-98, the large volume manufacturing potential of this approach is apparent.
Intel has announced they will offer flash memory products in 1997 using the Tessera format. An example of the Intel package is shown in Figure 10-99. The ball pitch is 0.75mm (30mils). The footprint is 5 rows by 8 rows, with a total of 40 balls. The Intel flash memory product will be the first large volume design win for the Tessera technology. Intel is expected to be only the first of many flash vendors to introduce products with this CSP approach. AMD and TI have stated their intent to use Tessera as well.

**CSP Barriers: Cost**

The chief barrier to widespread use of CSP are inertia, cost, and the requirements of the substrate. Figure 10-100 illustrates the rough cost of a CSP, at 2¢-7¢/pin compared with an SOP at 1¢/pin. It is expected that over time, and with volume production driving the learning curve, the cost of CSPs, especially using the Tessera approach, can reach the 1¢/pin range.
Figure 10-94. Compliance of Tessera CSP

Figure 10-95. Amkor's Chip Scale BGA (CSBGA)
CSP Barriers: Substrate Requirements

Though it may be straightforward to place a micro-BGA on a circuit board surface, being able to connect the solder balls to routing channels may be a challenge. With one track technology, and limited to just the surface routing layers, two rows of balls of a micro-BGA can be captured and routed per side. With two track technology, 3 rows deep can be captured from each side in one layer of interconnect. If it were possible to place small enough vias in the board to allow contact from a second layer of interconnect, the number of rows would double.
Figure 10-98. µBGA Assemblies on 70mm Reels

Figure 10-99. Intel's µBGA Flash Memory Technology
The number of I/O that can be routed from a CSP based on the number of rows deep that can escape is given by:

\[ N = \frac{4L}{P_s} \cdot N_{\text{rows}} - 4 \cdot N_{\text{rows}}^2 \]

This is shown graphically in Figure 10-101.

**Figure 10-100. CSP Pricing Relative to SOP & Flip Chip**

**Figure 10-101. Pin Count Limited by Routability for 0.75mm (30mil) Ball Grid**
For a grid pitch, \( P_g \), the solder pad size on the surface of the substrate is roughly \( \frac{1}{2} P_g \). This is also equal to the space that exists on the board between the solder pads to pass surface traces, and is the width of the routing channel. To fit one track technology within this gap, with a space on both sides of the line equal to its line width, \( w_1 \), requires a width of:

\[
w_1 = \frac{1}{3} \cdot \frac{1}{2} P_g
\]

For two track technology, the width, \( w_2 \), is given by:

\[
w_2 = \frac{1}{5} \cdot \frac{1}{2} P_g
\]

For the various micro-BGA and BGA pitches, the linewidths required for one and two track technology are shown in Figure 10-102. For example, one track technology with a 0.75mm CSP would require linewidths of 0.125mm or 5mils. Two track technology would require a linewidth of 0.075mm or 3mils.

This means that using 0.75mm grid pitch CSPs, a conventional 5mil line and space can route a two row deep CSP. A 10mm CSP can have about 100 balls. It is likely that most board shops will be able to fabricate 3mil line and space for a premium. This will allow two tracks, and a 10 mm CSP will be able to support about 120 I/O. This illustrates that using conventional printed circuit
board technology, CSPs will be limited to relatively low I/O count. Even going to a length of 15mm, two tracks would support only 200 I/O. All the proposed CSP applications, from Sharp, Sony and Intel are for pin counts less than 200, so as to take advantage of conventional linewidths, while keeping board cost down.

To take advantage of CSP and flip chip with more I/O requires a substrate with much finer linewidths and via pitches. One popular approach takes advantage of build up multilayer (BUM) technology for laminate substrate. This is reviewed in Chapter 11. A technology such as BUM will be a critical link for CSPs to expand in applications with higher ball counts than about 200.