EEE 6392 Microsystem Packaging

Examination 2008 - Model answers

Gavin Williams and Mark Hopkinson

Q1

a. Describe the methods of (a) wire bonding, (b) tape automated bonding, and (c) flip chip technology for the connection of circuits to a substrate or direct to a chip package. What are the advantages and disadvantages of these techniques?

Answer to include:

Wire bonding: Single gold wire bond to pins in lead frame.

Advantages

- Simple process
- Relatively inexpensive lead frame and tooling
- Easy to inspect and test (wire pull)
- Good for low volume, one-off production

Disadvantages

- Slow
- Difficult for small packages/large lead density
- Can suffer mechanical, thermal, oxidation damage (stress)
- Lower reliability (every lead different). Long term reliability problems

Tape Automated bonding: Multiple 'thermosonic' bonding to copper plate based 'pre-form' on tape.

Advantages

- Smaller bond pad & spacing available compared to wire bonded
- Increase in production rate, lower labour costs
- Less variations in bond geometry
- Stronger and more uniform inner lead bonding strength
- Better electrical performance, noise and frequency, higher pin outs and lighter weight.
- Greater lead density
- Chip can be attached in a face-up or face-down configuration

Disadvantages

- IC specific. Each die must have its own tape patterned for its bonding configuration
- Time and cost of designing and fabricating the tape
- Tooling (capital) expense of the TAB bonding equipment and skilled set-up

Flip chip: Solder bump on side, receptor on other. Chip flipped down onto package and solder re-flow to make connection

Advantages

- High density (short spacing, 2D array)
- Less expensive than other methods where package density is very high and volume high.
- Short interconnection length, lower parasitics. Can be compatible with high frequency
- Stronger and very uniform solder bumps possible
- Greater lead density
- Chip can be attached in a face (i.e. active surface)-up or face-down configuration

Disadvantages

- High tooling cost
- IC specific
- Complex metallization schemes, more lithography steps
- Can suffer from reliability issues; electromigration, corrosion.
- Difficult to inspect

b. The Intel 80286 microprocessor (made in 1995) had 10^5 gates and operated at 8MHz. 10 years later the first Intel Pentium (80586) processors had $2x10^6$ gates and operated at 100MHz.

(i) Assuming Rents rule, with the rent exponent =0.5 and the rent multiplier =0.2, calculate the possible number of I/O terminals for these two packages

(ii) The 80286 used a plastic leaded chip carrier (PLCC) package, whilst the 80586 used a ceramic pin grid array (PGA) package. The package area is 2cmx 2cm in each case. Estimate the average pin separation for the two types of package

(iii) Explain the advantage of the PGA and why the material of the package also had to be changed

(iv) The latest Intel Core 2 Quad processors are multichip modules consisting of 2 individual die each with 2 processors. Describe what is meant by a multichip module and why it is used instead of individual packaged processors

Terminals $= k.G^p$, where k= Rent multiplier, p= Rent exponent and G=no. logic gates

80286 Terminals = 63

80586 Terminals = 282

These numbers are slightly out with reality, but this does not compromise the question.

2 cm^2 package

PLCC 63 pin outs around edges, means about 16 per side. In 2 cm, means pin separation of 1.25 mm

PGA has 282 pins in 2D array. SQR(282) per line= 16.7.pin separation 1.2 mm

2D array of PGA can accommodate 4.5x more pins with lead similar spacing.

Due to higher number of logic gates and much faster clock speed, there is much more power dissipation in the 80586. This is somewhat offset by CMOS gate capacitance reductions and lower operating voltage. The plastic package could not cope with this increased power dissipation due to its poor thermal conductivity and susceptibility to damage (softening, melting). PGA very good in these aspects and has the mechanical stability to retain the pins.

Multi-chip module consists of two or more ICs within a single package and on a custom designed substrate. To operate as individual packaged processors the system would be compromised by off chip (i.e. package, PCB) parasitics and the limitations of the PCB bus speed. The MCM is designed to operate at the on chip speed and has a custom designed internal interconnect structure to do so. It removes the 'packaging bottleneck' relating to on-chip to off-chip issues. Individual packaged processors would have to operate at reduced speed and would have much higher power dissipation.

c. Describe the main mechanisms of signal degradation for the packaging of high speed integrated circuits. What steps should be taken in packaging design to minimise these effects?

Answer to include a description of:

Crosstalk

Simultaneous switching noise

Timing noise

EMI/EMC

All controlled by capacitive and inductance coupling and mismatch.

Conductors, such as leads and band pads with low sheet resistances. Low series resistance interconnects, short lengths. Low mutual inductance and coupling capacitance between interconnects, packaging insulating materials with low dielectric constants, reductions in parasitic inductances or capacitances through appropriate package design and modeling, impedance control and matching of the chip with the package and with the package with the with the printed circuit board.

d. One of the major sources of parasitic impedance introduced by a package is the mutual inductance between bond wires. The mutual inductance (M) of two conductors is given by:

$$M = \frac{\mu_0 l}{2\pi}\left[\ln\left(\frac{l}{d}+\sqrt{1+\left(\frac{l}{d}\right)^2}\right)-\left(\sqrt{1+\left(\frac{d}{l}\right)^2}\right)+\left(\frac{d}{l}\right)\right]$$

Where l is the conductor length and d the distance between conductors.

A typical Quad flat pack (QFP) package has leads of 6mm length with a separation of 1.2 mm. Calculate the mutual inductance between two leads.

What other sources of parasitic inductance could be introduced by the packaging?

Describe one modern package which effectively deals with the problem of mutual impedance between leads.

l/d= 5, d/l=0.2

$$M = \frac{\mu_0 l}{2\pi}\left[\ln\left(\frac{l}{d}+\sqrt{1+\left(\frac{l}{d}\right)^2}\right)-\sqrt{1+\left(\frac{d}{l}\right)^2}+\frac{d}{l}\right]$$

=μo.l/2.pi x (2.3-1.02 +0.2)

=μo.l/2.pi x 1.48

4.pi x1E-7x 0.2/2x 3.142=4E-8

4E-8x 1.48= 59nH – this value is a bit high, but it does not compromise the question.

Inductive or capacitive ground plane coupling. Impedance mismatch.

BGA (Ball grid array) is very effective due to 2D layout and short interconnection length

Q2

a. Explain why thermal management is important for packaged CMOS integrated circuits. What is responsible for the heat generation?

Describe the problems which could result if the temperature is not managed?

What general methods are there for providing heat flow from a package?

In CMOS each logic gate change involves the charge/discharge of a gate capacitor and there is also gate leakage. Heat is also dissipated due to parasitic impedance losses in the chip or from the chip to the package. There is a very high heat density involved. There is also significant resistance to heat flow is presented by the bonding material between the chip and substrate, the substrate itself, and the thermal

conducting path to the exterior surface of the package or the PWB upon which the package is mounted.

Without thermal management the maximum on-chip temperature would be reached.

Control temperature to prevent catastrophic thermal failure (damage to IC, damage to packaging, thermal stress, cracking delamination, melting etc. Also helps to extend the useful lifetime of the electronic system

Need to place packages in contact with solid, liquid or gas at lower temperature to provide heat flow away from the package.

Examples can include solid metal heat radiators, forced air cooling (e.g.: fans) or immersion in liquid which can vaporize

b. Under typical operating conditions an integrated circuit of dimensions 5mm x 5mm, dissipates 10 W. The IC is epoxy mounted to a 3 mm thick Silicon Nitride substrate, which is then epoxy mounted to a 10 mm thick aluminium heat sink. The epoxy is spread to a thickness of 100 μm.

The maximum allowable on chip temperature is 95^{o}C. Assuming linear heat conduction, what is the maximum allowable temperature on top of the heat sink? What methods could be employed to keep this temperature below the limit?

$R_{th} = L/k.A$ (°C/W) and $\Delta T = Q \, . \, R_{th}$

$k_{epoxy} = 2W.m^{-1}.K^{-1}$, $k_{SiN} = 65W.m^{-1}.K^{-1}$, $k_{aluminium} = 216W.m^{-1}.K^{-1}$

Area= 2.5E-5m^2

R_{th} Epoxy= 2, SiN=1.85, Al=1.85

So, from IC to surface, the temperature change, T = 95-(10x2)- 10x1.85 -(10x1.85)-25x1.85)

T on Al surface = 18^{o}C

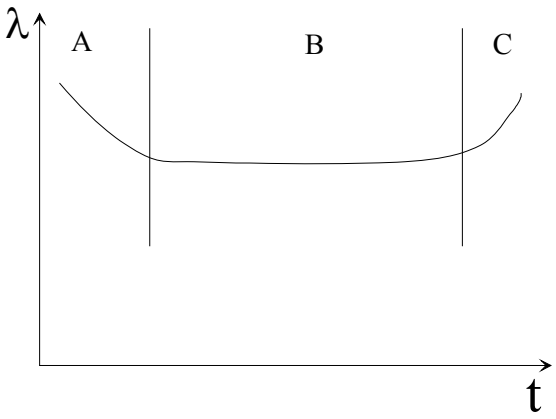
Need to use forced air cooling of large finned heat sink

c. Integrated circuits show an average failure rate as a function of time which is typical of semiconductor devices.

Draw a graph of the average failure rate versus time and describe its three main regions.

Describe the process of 'burn in', which could be applied to these devices. Show a typical failure rate versus time graph for an IC during 'burn in'.

What advantages might this have for the manufacturer?



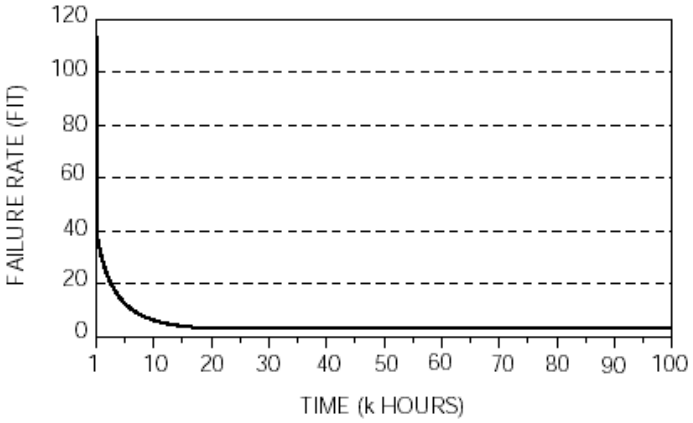
A. High "early failures" or "Infant Mortality" due to manufacturing defects

B. "Midlife" or "Steady state" period of low and generally constant failure rate

C. "Final" or "wear out" period

Burn-in: Manufactured devices often show early failure. Generally, manufacturing defects cause the infant mortality failures ICs: pinholes, photoresist or etching defects resulting in near-opens or shorts. Package: scratches, cracks, weak chip mounts or wire bonds. The purpose of the burn-in procedure is to operate the devices for some time during which most of the devices that are subject to Infant mortality failure actually fail

The conditions during burn-in accelerate the failure mechanisms that contribute to infant mortality failure



For a manufacturer it reduces the number of returns of failed devices. Allows them to demonstrate greater reliability

d. Describe what is meant by the 'accelerated testing' of semiconductor devices and why is it performed? Which typical environmental stresses are used to accelerate failures?

HAST is a common industry standard acceleration method. Describe its parameters. What failures could it promote?

A semiconductor laser device should achieve at least $1x10^5$ hours of reliable operation at its 20^oC (293K) operating temperature. However only 1000 hours are realistically available for testing. Assuming the failure mechanisms are known to be due to a thermally activated process with a rate of failure given by;

$R = R_0 \exp(-E_a / k_B T)$,

where R is the rate, E_a the activation energy, k_B is Boltzmann's constant and T the temperature. At what temperature should the accelerated testing be performed?

Assume that E_a = 0.5eV

Accelerated Testing: Impossible to test the required reliability under normal operating conditions. This necessitates a means to accelerate the mechanisms that cause devices to fail

- temperature
- voltage
- current
- temperature cycling to accelerate mechanical failure of chips and assembly package
- Humidity
- Mechanical stress/vibration etc

Highly Accelerated Steam and Temperature (HAST). Same basic function as 85/85, in typically 10% of the time, making HAST tests useful for immediate feedback and corrective action.

Test Conditions: 120 °C, 85% RH, biased,100 hrs.

Test typically identifies assembly quality.

Test Used: Temperature Cycling

Test Conditions: -65°C to +150°C, 1000 cycles

Promotes failures due to poor hemeticity, leading to water vapour intake and then to aspects such as metallic corrosion. Induces stress on bonds, lead wires.

For failure rate

$$R = R_0 \exp(-E_a / kBT)$$

Two failure rates $$\frac{t_1}{t_2} = \frac{R_2}{R_1} = \exp\left[\frac{E_a}{k_B}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right]$$

T1 = 20 ^{o}C= 293 K, T2=?, t1=1E5, t2=1E3

E_A = 0.5eV = 8E-20J

E_A/K = 5797

1E2 = exp ()

ln(1E2) = 5797($1/T_1$-$1/T_2$)

($1/T_1$-$1/T_2$) = 7.94E-4

T2 = 1/(1/293-7.94E-4)

= 381 K or 108 ^{o}C

Q 3

a. Describe the soldering method by which surface mount components are attached to the top surface of a printed circuit board (PCB).

Soldering method = reflow soldering (*not* wave soldering)

Details to include:

- Screen printing of solder paste onto pads on PCB
- Solder prior-deposited onto component legs
- Placement of components onto pads
- Reflow soldering. 4 stages:

Pre-heat

Slow* ramp (~1 °C/s) from room temperature up to T ~ T_m - 10 °C

(* to avoid thermal shock)

Evaporation of solvent to leave weakly coalesced solder powder and flux

Removal of moisture from package in order to avoid delamination

Soak

Hold at T ~ T_m - 10 °C

Allow different thermal mass components to become isothermal

Flux activated, leading to removal of surface oxides

Reflow

T ~ T_m + 30 °C for a short time

Solder particles melt and 'reflow'. Surface tension causes them coat the pad and component legs.

Cool

Solder solidifies T < T_m

Cooling rate as fast as possible without causing thermal shock

Slow cooling can lead to formation of large (brittle) solder grains

Fans often used to enhance cooling rate

b. The packing density can be increased by mounting components on both side of the PCB. What are the limitations of this approach and how could they be overcome?

Limitations: gravity causes heavy components to fall off the underside of the board during the preheat stage.

Overcome by only putting heavy components on top surface. Use tacking compound if necessary to put heavy components on underside.

c. An extra assembly step usually follows the soldering of ball grid array (BGA) packages onto the PCB. Describe this step and explain its purpose.

Underfill to help with stress relief. Deposit either as a solderable pad or by capillary flow of polymer after soldering.

d. The tin-lead solder that was used in electronics manufacture had a eutectic composition, whereas the many different 'SAC' alloys that have replaced it are typically not of a eutectic composition. Discuss the issues relating to this switch, with special regard to the manufacturing process and subsequent product reliability.

Non-eutectic composition, therefore non-abrupt melting/solidification temperature, hence formation of potentially brittle inter-metallic phases. Also SAC (Sn (tin) – Ag (silver) – Cu (copper)) solders have higher T_m (~ 200 °C) than Sn-Pb, hence higher energy costs. Also more expensive. Many different compositions, therefore possibility of conflict with composition of solder on component legs.

Q 4

a. An electronic system includes two BGA-packaged devices - a high speed microprocessor and a video-co-processor. They are interconnected via a 32-bit data bus. Discuss the design, placement and routing of these two devices on a PCB and the overall distribution of power, ground and signal tracks.

PCB design: Keep north-south tracks on one layer, east-west on another. Choose design rules (minimum track and gap, via size) that will allow routing from the fine-pitch BGAs.

PCB placement: Place components close to one another and in-line in order to simplify routing, but ensure adequate thermal cooling, since this configuration may lead to hot spots on the PCB.

PCB routing: high speed signals therefore use stripline (i.e. ground planes above and below signal layer). Minimise the number of bends in tracks and vias.

Power and ground: continuous sheets where possible in order to minimise inductance for high speed signals and to act as efficient thermal conductor.

b. The BGA package has a thermal resistance of 0.3 °C/W and the video co-processor dissipates 8 W. What is the maximum thermal resistance of the heat sink that could be used to maintain the junction temperature of the device below 125 °C?

$R_{BGA} = 0.3$ °C/W $Q_{VCP} = 8$ W $T_{max} = 125$ °C $T_{room} = 20$ °C

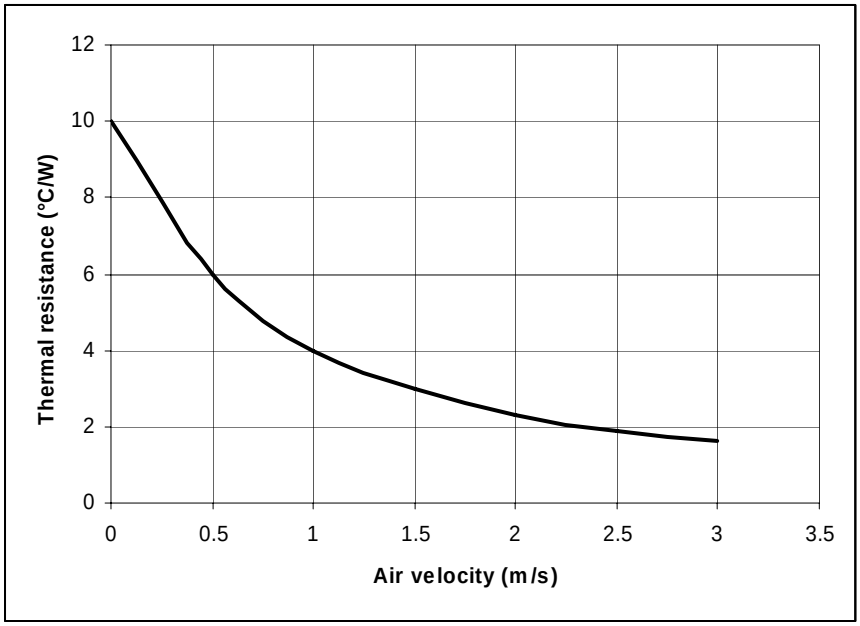
From Fourier's Law of heat flow: $\Delta T = QR$

$$\Delta T = T_{max} - T_{room}$$

$$R = R_{junction\text{-}case} + R_{case\text{-}ambient} = R_{BGA} + R_{heat\ sink}$$

Therefore: $R_{heat\ sink} = (T_{max} - T_{room})/Q_{VCP} - R_{BGA} = (125\text{-}20)/8 - 0.3 = 12.8$ °C/W

c. The microprocessor dissipates 45 W and is mounted with a heat sink cooled by a fan. Using the graph below determine the fan speed necessary to keep the chip within a safe working temperature range.



Thermal performance of an air-cooled heat sink

P = 45 W T_{max} = 125 °C

As above: $R_{heat\ sink} = (T_{max} - T_{room})/P_{VCP} - R_{BGA} = (125-20)/45 - 0.3 = 2$ °C/W

Hence from graph: Minimum required air velocity = 2.25 m/s

d. Describe the alternative cooling strategies that might be employed if it were desired to implement the electronic system described above as a portable (i.e. low power) item.

- Thermoelectric cooler

Based on Peltier effect: $\Delta T = V/\alpha$

Formed as an array of p-n junctions of semiconductor with low α

Voltage control allows temperature stabilisation.

Heat must be removed from the hot surface

- Heat pipe

Based on the high latent heat of evaporation of liquids

Formed as a sealed tube containing a small volume of liquid (water, alcohol)

Inside surface of tube has a wicking structure (small grooves or fine mesh)

Liquid boils at hot end of tube (thus absorbing heat) and condenses at cool end (thus releasing heat). Capillary action draws liquid back to hot end in order for process to continue.