Cost-minimized Double Die DRAM Packaging for Ultra-High Performance DDR3 and DDR4 Multi-Rank Server DIMMs

Richard Crisp[1], Bill Gervasi[2], Wael Zohni[1], Bel Haba[3]
[1]Invensas Corp, 2902 Orchard Parkway, San Jose, CA USA
[2]Discobolus Designs, 22 Foliate Way, Ladera Ranch, CA USA
[3]Tessera Inc, 3025 Orchard Parkway, San Jose, CA USA

[1]E-mail: rcrisp@invensas.com

Abstract
A cost-minimized generation-spanning double die DRAM packaging (DDP) technology suitable for making ultra-high performance high-capacity server DIMMs for both the DDR3 and DDR4 DRAM generations was developed. Using existing wirebond-based manufacturing infrastructure it is immediately deployable with no new assembly equipment required.

Significant results were obtained relating to performance enhancement and cost reduction versus existing packaging and DIMM designs. Both die are mounted face down in the package with each showing identical performance. Bin split yields are enhanced significantly. The package ballout features placement of the command and address terminals in the center of the package permitting single-layer routing of the timing-critical address and command bus at the DIMM level. Data and Data Strobe signals have shorter routing stubs on the DIMM. The improved PCB layout resulting from the new ballout resulted in a quadrank RDIMM capable of operating at over 1600MT/s in a two DIMM per channel configuration: 50% faster than standard DIMMs.

The face-down laterally displaced die arrangement with a sub-1mm package thickness reduces the thermal impedance by 25% versus conventional DDPs. Total assembly cost is the lowest of any DDP and on a per-die basis is lower than Single Die Packaging.

Keywords
DDR3, DDR4, RDIMM, LRDIMM, Hypercloud, DDP, wirebond, RDL

1. Introduction
The design approach used was to identify clock-rate limiting features of standard DRAM package and DIMM designs and to correct them using co-design of the DRAM package/ballout and DIMM while looking for opportunities for cost reduction. Focus areas for performance scaling were stub minimization for the DQ / DQ Strobe nets and command-address (C/A) bus, layer count reduction in the DIMM raw card and improved thermal performance.

Conventional double die DRAM packages (DDPs) are designed to be footprint-compatible with single die packages (SDPs). All feature die stacked atop each other without lateral displacement. Electrical connections are accomplished using wirebonds. The simplest type features back to back die with the lower die mounted face-down and the upper die mounted face up (Fig 1a). Data bus signals from the two die are wire-or connected. The upper die has long wirebonds that connect to the centrally-located bond pads and extend half way across the die to be bonded to the underlying substrate. The lower die wirebonds are made through a window in the substrate like conventional single die packages. The structure therefore has asymmetric signal and power delivery connections to the two die that causes the upper die's electrical performance to be degraded.

Another approach uses a redistribution layer (RDL) on each die to move the bonding pads from the middle spine to the periphery region and mounts both die face up with the two die separated by a spacer die so that the wirebonds from the lower die are not crushed by the upper die (Fig 1b). The RDL process adds significant cost, approximately $300 per 300mm wafer, and the added RDL structure increases the capacitance of the signals routed across the die, degrading the performance.

For both approaches, because the two die are stacked vertically without lateral displacement, some of the heat from the lower die must pass through the upper die, increasing its operating temperature. Furthermore both schemes feature topside wire loops that must be protected by overmolding. That increases the thickness of the package, further degrading the thermal performance versus single die packages.

2. Package Structure
The double die package developed in this work overcomes each of the shortcomings of the previous DDPs described above. Figure 2 shows the new structure. Each die is mounted face down and is wirebonded through its own window. We have called this structure Dual Face Down (DFD™). Electrically both die perform near-identically as shown in the shmoo plot of Figure 3. The data was measured on a production VLSI tester from a sample of 15 randomly chosen units operated at 95C.

The DFD has no top-facing wire loops so it needs less topside encapsulation resulting in a thinner structure of less than 1mm. Combined with the lateral displacement of the two die, the thermal impedance is reduced by 25%

versus conventional DDPs. (Figure 4). The thinner package also reduces airflow resistance between DIMMs installed in systems further improving the cooling efficiency.

Because both die have nearly identical electrical performance matching that of an SDP DRAM, bin split yields are significantly improved over standard DDPs. Figure 5 shows production sort results for a controlled experiment wherein 1000 units of the DFD were compared with 1000 control units. The control units were designed to mimic the top die of the opposing face DDP of figure 1a. Die for the experiment came from the same wafer lot.

Since the control did not include substrate routing to connect to the lower die nor did it include the lower die, the sort results for the control are somewhat optimistic. Yet each die in the DFD showed a 67% improvement of sort yield to the highest speed bin versus the control. Equally important, no difference was observed in top versus bottom die sort yield indicating truly symmetric performance of the two die; monolithic performance in a double die package was achieved.

The DFD has a simple and short process flow through the manufacturing line. It is the only DDP that can be made using single pass wirebonding. Manufacturing the DFD therefore costs less than either of the two DDP alternatives of Figure 1.

A detailed assembly cost model revealed that the packaging cost per die assembled using the DFD structure was lower than either of the DDPs and is actually less than the SDP due to fewer process steps required for packaging a given number of die (Table 1).

The cost of gold bond wire dominates the cost of the opposing face DDP of Figure 1a, while RDL cost dominates the dual face up DDP of Figure 1b. For example, assuming $300 RDL cost and 1200 good die per wafer, the RDL cost alone is $0.25/die, which is 25% more than the entire packaging cost of a SDP DRAM according to the cost model.

An added benefit of the fact that the DFD represents the lowest cost method for packaging a given number of DRAM die is that DIMMs can be made using single sided assembly without compromising capacity, further reducing manufacturing cost for a given capacity DIMM: they are not limited to quadrank usage only.

3. Ballout Considerations

The ballout for the DFD was developed in conjunction with the design of a quadrank DDR3 RDIMM. The critical nets that were optimized were the data / data strobe nets and the C/A bus.

Double sided assembly used for RDIMMs requires that like-named C/A bus signals for pairs of packages placed in discrete sites on opposite sides of the DIMM from one another are electrically interconnected forming localized breakout regions. A global C/A bus routes through each breakout region to connect each site to the common C/A bus (Figure 6).

Connections from the breakouts to the common C/A bus create stubs that can be up to 7mm long (Figure 7). At low frequency operation the stubs have little effect on the signal integrity, but as frequencies increase, signal reflections degrade the signal quality and timing margin and can cause functional failure. This creates a serious obstacle for operation beyond 1600MT/s.

Mass-market DRAMs have always had signal terminals arranged into two groups that are laterally displaced on opposite sides of the centerline of the DRAM die with a gap between them. The first DRAMs were packaged in Dual Inline Packages (DIP). The DIP was replaced by the SOJ followed by the TSOP. In each case the signal terminals were disposed in single columns placed on opposite edges of the DRAM package.
When the chipscale BGA was introduced, the single column of terminals evolved into two groups of three columns of terminals separated by a window through which wirebonds that interconnect the DRAM pads to the package substrate are placed.

In all cases there was a gap between these two groups of signal terminals placed on opposite sides of the centerline of the DRAM die (Fig 8). This is the primary factor determining the length of the stubs on the C/A bus in the breakout region. Only recently has the length of the stubs become a limiting factor for DIMM frequency scaling.

Exploiting the fact that the DFD has two laterally-displaced bonding windows (Fig 2) on the substrate it is practical to place the C/A bus signals in the center of the package in adjacent columns (Fig 9). This greatly simplifies the cross-tie routing in the breakout region leading to a significant reduction in C/A bus stub length to 2mm for each signal in the breakout (Fig 10). A simulation comparing the C/A signals on DIMMs constructed using conventional DRAM packaging versus the DFD is shown in Fig 11. The simulation shows a 16% improvement in voltage margin at 75% higher operating frequency for the DFD. The signals for the DFD also show significantly less ringing versus the conventional ballout. The simplified breakout routing allows the outer surface of the DIMM PCB to have a higher percentage covered with a solid copper flood, leading to improved heat spreading and a more effective power plane with reduced signal impedance discontinuities arising from breaks in the reference plane.

The signal ordering for the DFD's C/A terminals was set so that it mates without signal crossings to the register on the RDIMM used to buffer the C/A bus. This signal assignment scheme allows wider signal spacing reducing crosstalk. The combination of these advantages allows the C/A bus to be routed on a single layer (Fig 12)

versus the two layers required when a conventional ballout is used. Using a single layer for routing timing-critical C/A bus signals avoids differential propagation velocity and impedance variation of the signals when carried on separate layers, both of which degrade timing margin. Note that this timing skew cannot be compensated by the memory controller unlike data bus skew. As operating frequency is increased, this advantage will become of greater significance.

The DQ and DQ Strobe nets are timing-critical and are sensitive to stub lengths. In order to minimize the length of the stubs that interconnect the various DQ and DQ Strobe signals together, the terminals were placed adjacent to the top edge of the package as shown in the ballout of Fig 9. By orienting the DFD packages that are placed on the same side of the DIMM PCB in a configuration so that the DQ/DQ Strobes from a pair of such packages are facing each other, the length of the stub interconnecting the signals is minimized as shown in Fig 12. This reduces the stub length to 4.5mm compared to a conventional BGA ballout at 6.5mm.

4. Substrate Design

Using a four layer substrate in the DFD package allowed the DQ nets inside the DFD package to be referenced to the Vss plane with the Address and Clock nets referenced to the Vdd plane (Fig 13) as specified in the JEDEC DIMM standards for the DIMM PCB. This further improves the signal integrity by implementing a consistent set of design rules for signal return current paths for the entire system signal path from memory controller through DRAM package to the DRAM.

5. Applicability to LRDIMM and Hypercloud™ DIMMs

The LRDIMM differs from the RDIMM in that the DQ and DQ Strobe signals are buffered[1]. The data buffer is placed in the central region of the LRDIMM. This requires all data and data strobes to be routed from each DRAM package to the buffer and then routed back to the edge connector which demands additional routing layers versus an RDIMM. Since the LRDIMM is plugged into an edge connector, the thickness of the DIMM PCB is fixed.

Adding PCB layers necessarily requires a reduction of the thickness of the dielectric layers separating the power planes and routing layers. Unless the width of the traces is made narrower, the characteristic impedance of the etched traces is decreased and can lead to signal reflections arising from impedance discontinuities that diminish voltage and timing margin.

Trace width is limited by the precision of the control of the etching process, with such narrower traces being more costly to manufacture within tolerance. Because the DFD's C/A bus routes on a single layer and other interconnections lay out cleanly, the layer count is reduced leading to nominal impedances being attainable with normal dimensional control keeping raw card costs from rising.

The Hypercloud architecture is similar to the LRDIMM in that the DQ and DQ Strobe signals are buffered, but unlike the LRDIMM the buffering is provided by a number of data buffer devices placed between the edge connector and the DRAM package array on the DIMM PCB[2]. The 11.5 x 11.5 mm package outline of the DFD supports placement of the buffers without requiring growth of the vertical height of the DIMM. In fact a simple modification of the RDIMM PCB will enable the Hypercloud data buffers to be mounted on the PCB making conversion of an RDIMM design to Hypercloud a straightforward matter.

6. Measured Results

A quadrank RDIMM was produced using 72 one gigabit die placed into 36 DFD packages (Fig 14). Unlike RDIMMs based on JEDEC-approved raw cards which are limited to 800 to 1066MT/s operation when used with two DIMMs per channel (2DPC), the DFD-based RDIMM was demonstrated to operate over 1600MT/s using die that when offered in SDPs, are sold up to 1866MT/s.

An eye diagram measured by probing the DQ and DQ Strobe signals of an RDIMM using DFDs in a 2DPC-configured system operating at 1600MT/s is shown in Fig 15. Probes were soldered to the signals on the DRAM side of the series damping resistors and measurements were made using a 12GHz bandwidth Digital Signal Analyzer (DSA).

To validate the symmetry of the performance of the two die contained within a single DFD package, probes were soldered to the DQ and DQ Strobes of the two die contained within the same package. The DQ and DQ Strobe signal waveform transitions and voltage levels are nearly identical (Fig 16).

7. Applicability to Future DRAM Generations

The DDR4 generation of DRAM is targeted to reach system operational speed up to 3200MT/s, approximately twice the speed of the highest official DDR3 system speed today[3]. Reaching these speeds will require great care being given to signal integrity and PCB layout. Stub lengths, differential propagation velocity induced skew and noise on signal reference planes will need to be minimized. The DFD's characteristics offer advantages in solving all of these challenges.

The numerous improvements offered by the DFD are directly usable for dual die DDR4 packages with all the same advantages offered to the DDR3 DRAM generation providing a generation-spanning multi-die packaging technology. As data transfer rates scale to 3200MT/s and beyond the many advantages of the DFD and its ballout scheme will become of greater importance and may

represent the only feasible way to reach these system frequencies.

In particular the central location of the C/A terminals combined with single-layer routability may therefore prove to be an important innovation in DRAM packaging and become the preferred ballout scheme for future generations of DRAM or other high speed memory technologies that may replace DRAM.

The DFD overcomes the performance disadvantages of previous DDP packaging for DRAMs. These important performance advantages combined with the fact that the DFD is the lowest cost method for packaging a given quantity of DRAM may completely change the landscape of how memory die are packaged in performance demanding applications.

Table 1: Cost Summary.

Package Type	Total Assembly & Package Cost	Per Die Assembly & Package Cost
DFD	$0.34	$0.17
DDR-RDL	$0.83	$0.4175
DDP-Opposing Face	$0.42	$0.21
SDP	$0.20	$0.20

9. References

[1] http://www.edn.com/article/519386-Basics_of_LRDIMM.php

[2] http://www.netlist.com/products/ppt/HC_Gibabyte_Brief_1.0.pdf

[3] http://ddr4.org/

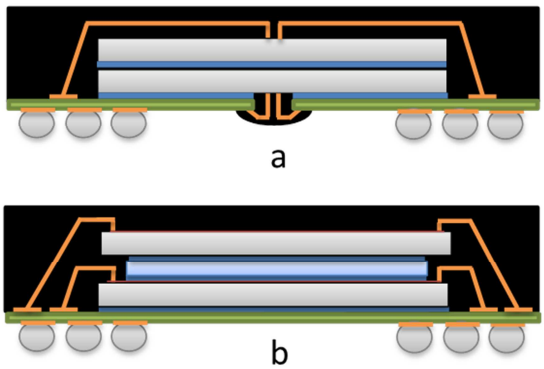


Figure 1: Conventional DDP packages

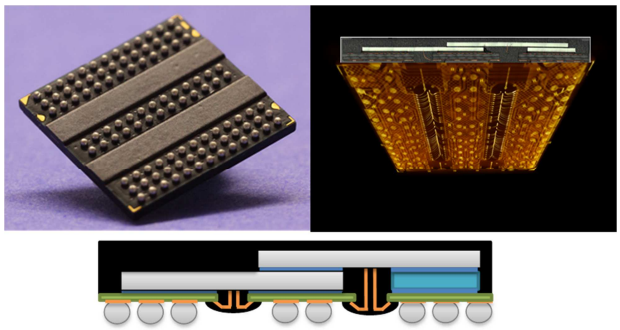


Figure 2: Dual face down package (DFD)

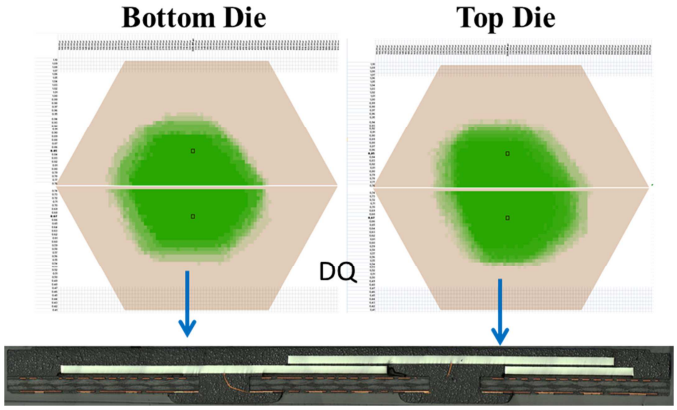


Figure 3: Shmoo plot from VLSI tester @ 95C, 15 unit sample size

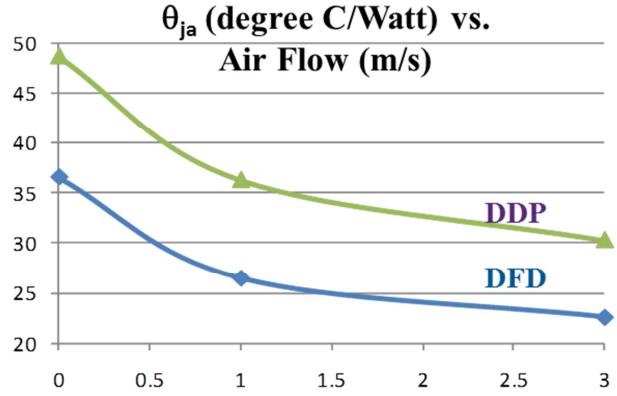


Figure 4: Thermal Impedance vs Airflow

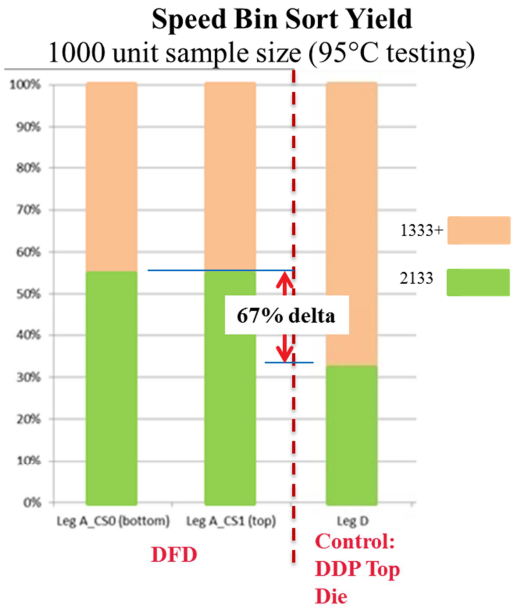


Figure 5: Production speed binning results

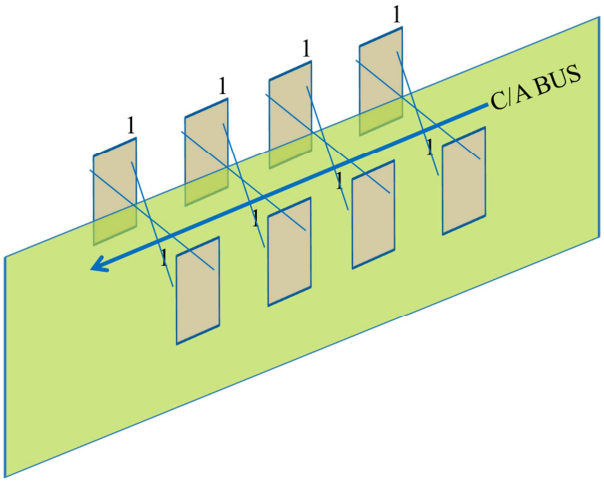


Figure 6: Double sided assembly electrical connections

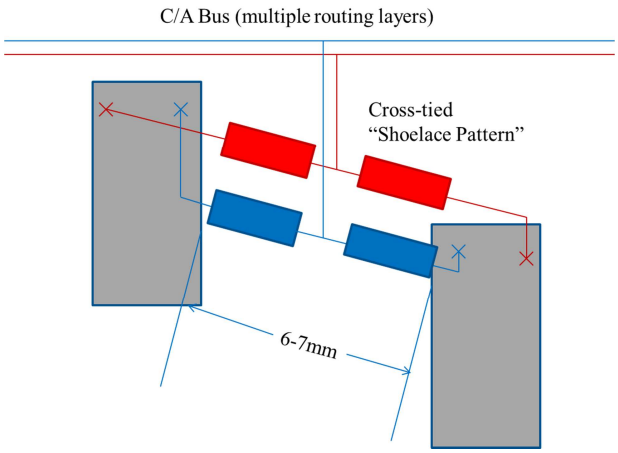


Figure 7: C/A bus stubs in breakout region

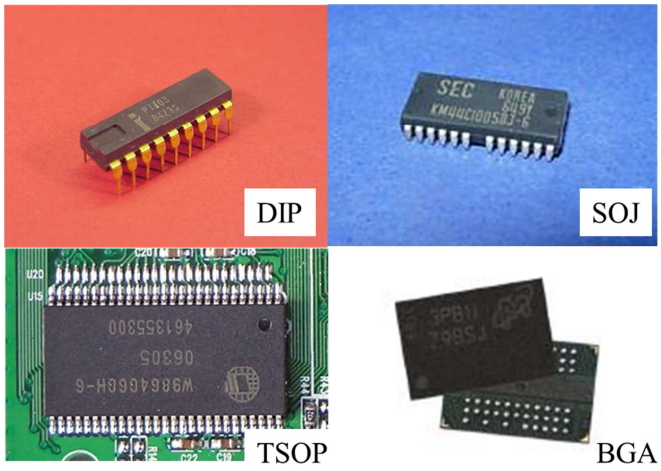


Figure 8: DRAM historical package evolution

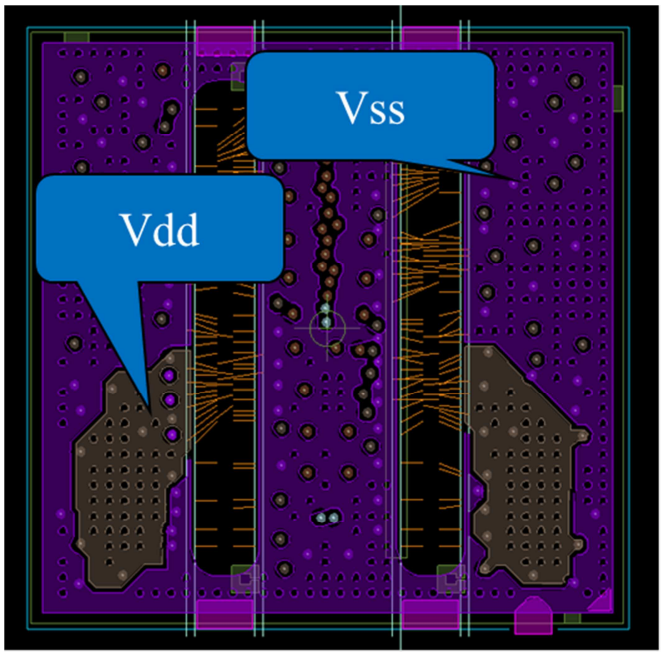


Figure 13: DFD substrate layout showing reference planes

	1	2	3				7	8				12	13	14
A	VDD	DM_1	VSS				VDD	VDD				VSS	DM_0	VDD
B	VSS	DQ0_1	DQ2_1				CK	CKB				DQ0_0	DQ2_0	VSS
C	VDD	DQ1_1	DQ3_1				RASB	CASB				DQ1_0	DQ3_0	VDD
D	VSS	DQSB_1	DQS_1				A10	WEB				DQS_0	DQSB_0	VSS
E	ZQ_1	VDD	VSS				A15	BA2				VSS	VDD	ZQ_0
F	VDD	VSS	VDD				BA0	A12				VDD	VSS	VDD
G	VSS	ODT_1	VSS				BA1	A0				VSS	ODT_0	VSS
H	VREFDQ_1	CKE_1	VDD				A3	A1				VDD	CKE_0	VREFDQ_0
J	VDD	CSB_1	VSS				A4	A2				VSS	CSB_0	VDD
K	VREFCA_1	VSS	VDD				A5	A11				VDD	VSS	VREFCA_0
L	VSS	VDD	VSS				A6	A9				VSS	VDD	VSS
M	VDD	VSS	VDD				A7	A14				VDD	VSS	VDD
N	VSS	VDD	RSTB_1				A8	A13				RSTB_0	VDD	VSS

Figure 9: DFD ball assignment

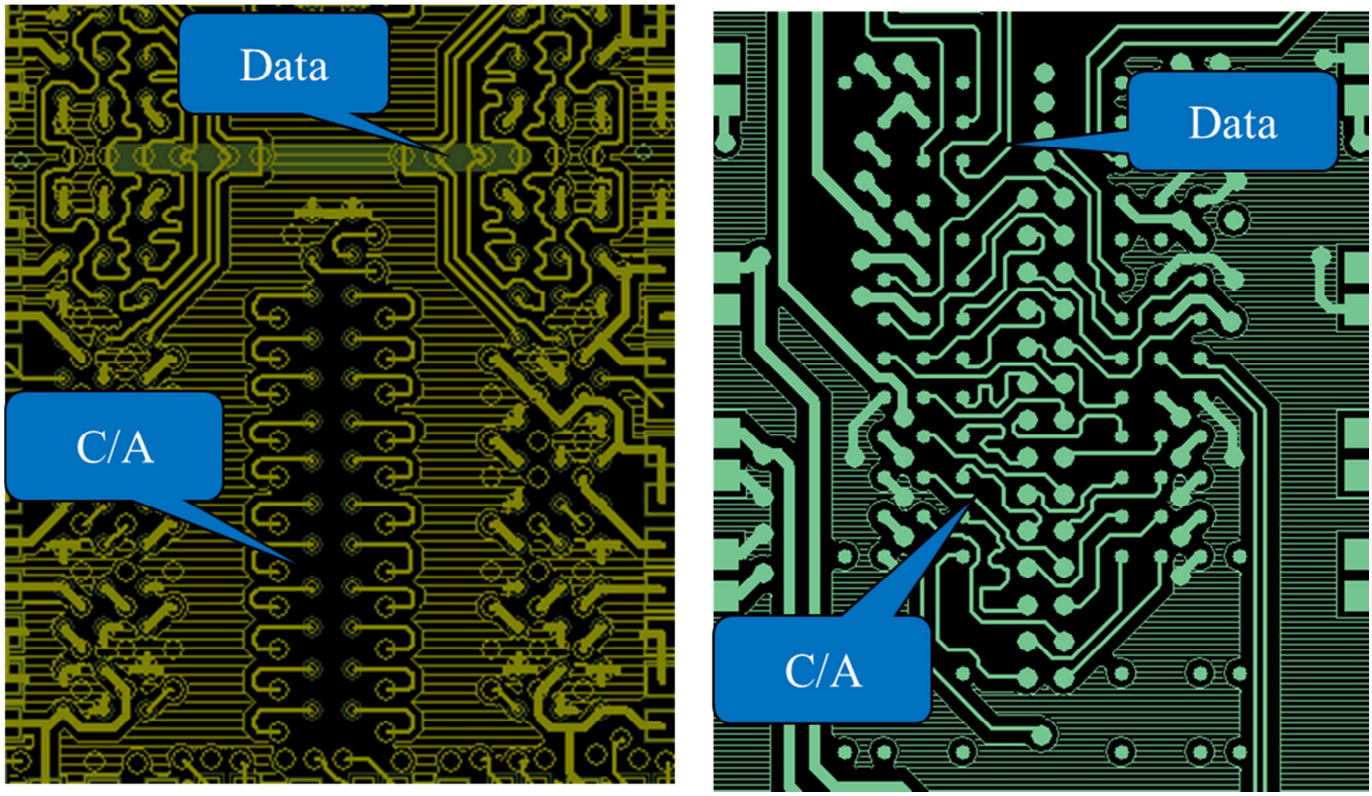


Figure 10: RDIMM breakout region detail: DFD (left), Conventional JEDEC (right)

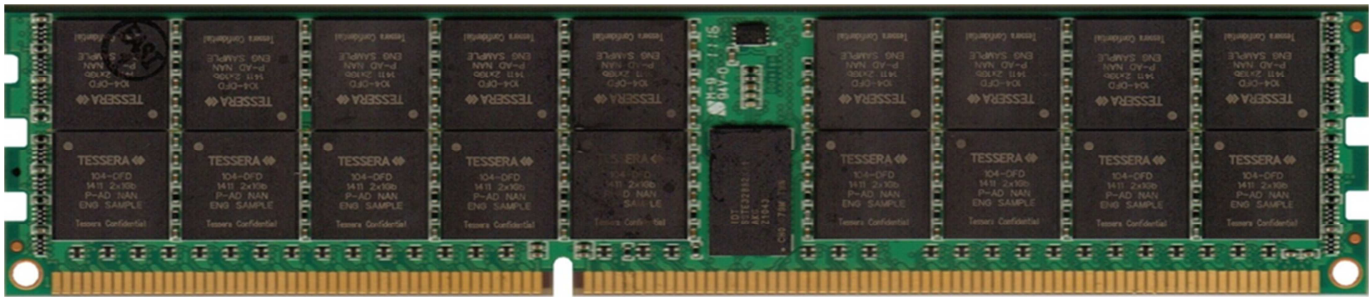
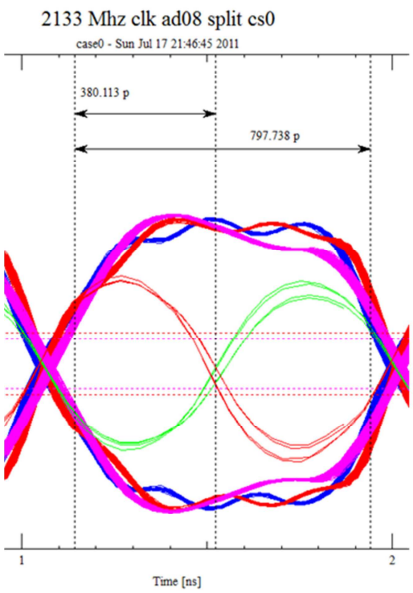
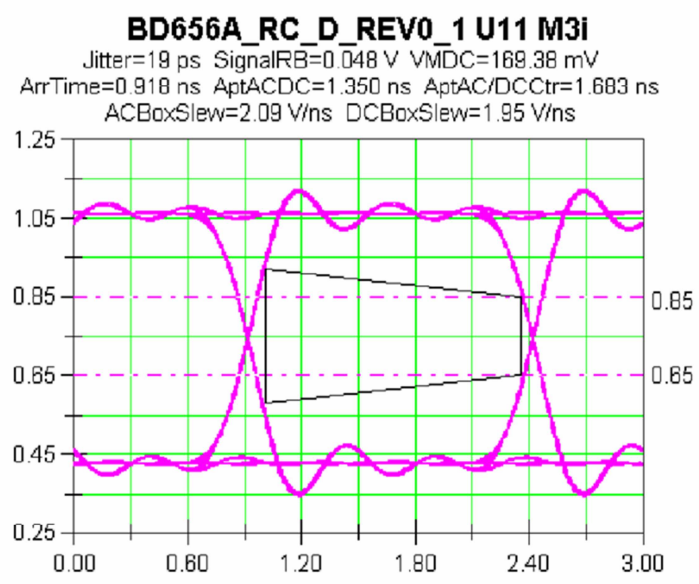


Figure 14: Quadrank DDR3 RDIMM using DFD packages





DFD @ 2133

Window = 798 ps

Ideal tCK = 938 ps

85% of a tCK

JEDEC r/c D @ 1333

Window = 1093 ps

Ideal tCK = 1500 ps

73% of a tCK

Figure 11: C/A bus simulation: DFD RDIMM vs conventional raw card

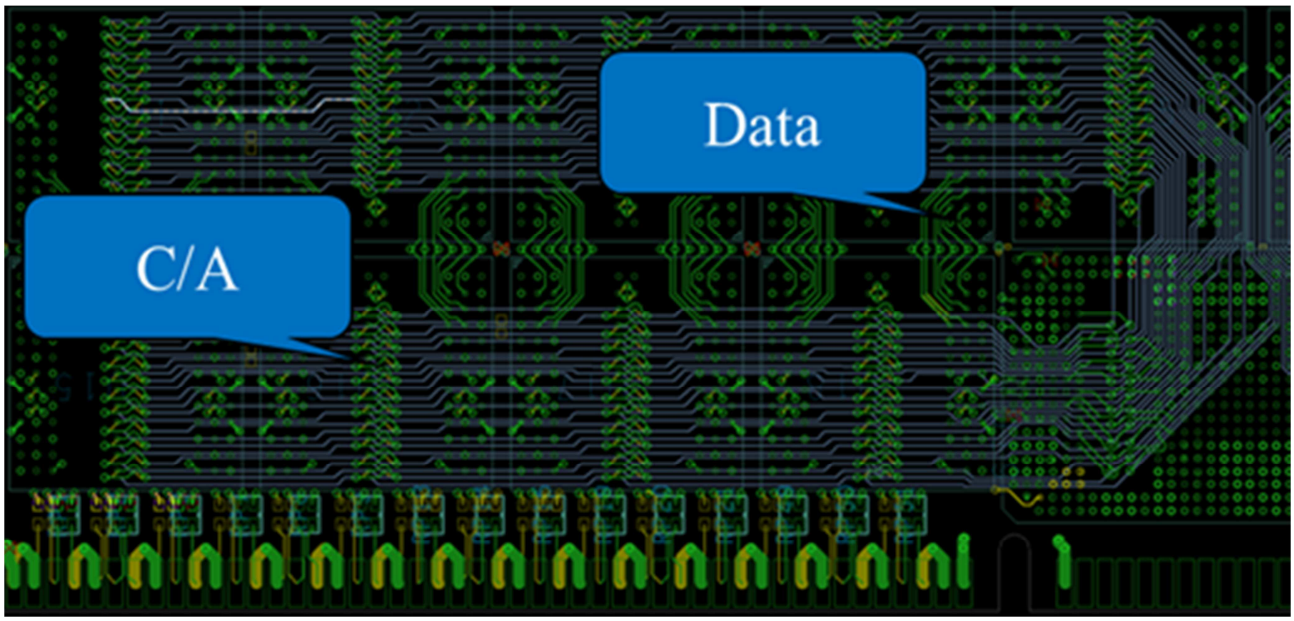


Figure 12: single layer C/A Bus routing through breakout regions

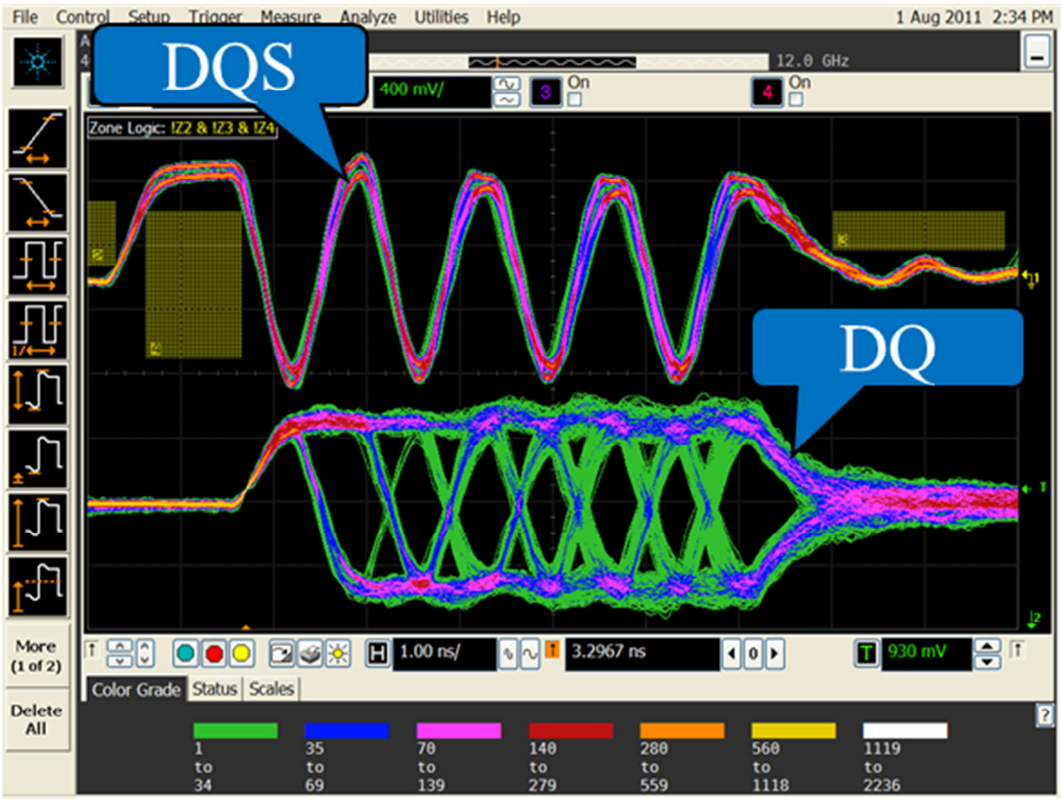


Figure 15: Read cycle eye diagram @ 1600MT/s

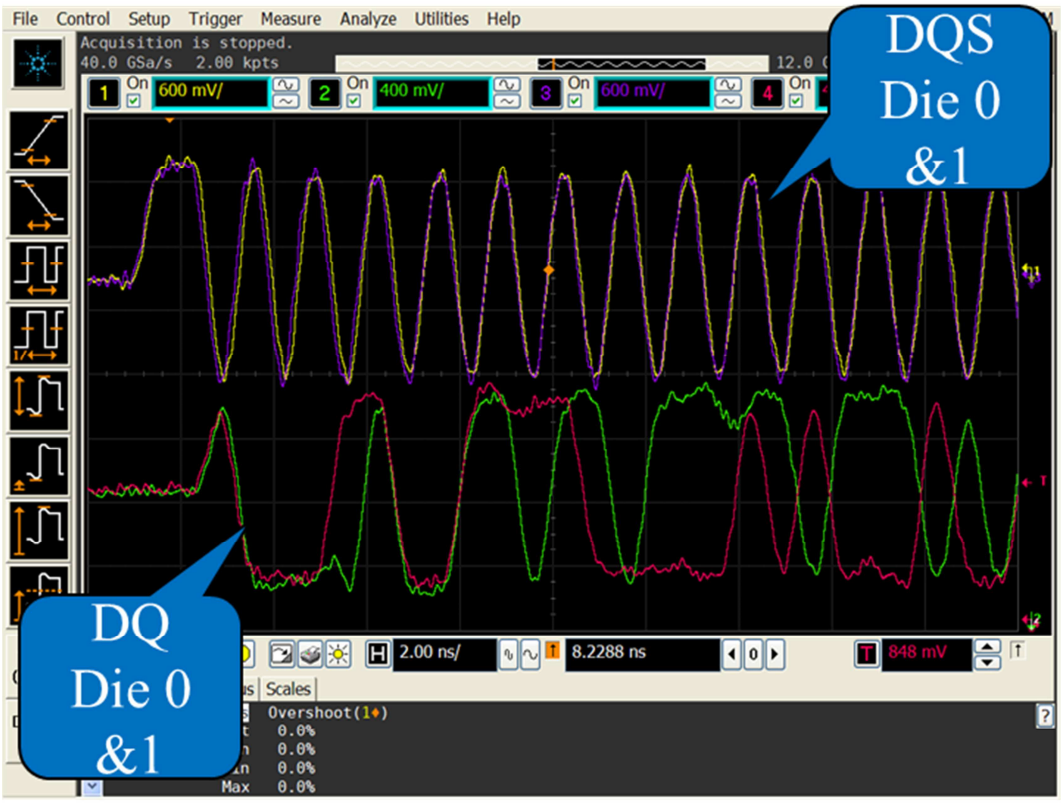


Figure 16: Read Cycle showing symmetric performance of both die in same package