High performance HBM
Known Good Stack Testing

Marc Loranger
FormFactor

John Oonk
Teradyne
Overview

- High Bandwidth Memory (HBM) Market and Technology
- Probing challenges
- Probe solution
- Power distribution challenges
- PDN design
- Simulation and measurement of final PDN design
- At speed test signal Integrity
- Summary
High Bandwidth Memory (HBM)

• **Market requirement**
  – Increase data bandwidth well above current GDDR5 technology
  – Decrease power per GB/s of bandwidth
  – Smaller size
    • Improve power distribution
    • Signal transmission

• **Long term roadmaps**
  – Expand into server applications and high performance computing when reliability is proven
High Bandwidth Memory (HBM)

• Next generation DRAM memory architecture
  – Four independent channel stack
  – Very wide data bus
    • 128 bits per channel
    • 512 bits total
  – Data bandwidth
    • HBM is up to 128GB/s per stack
    • GDDR5 is 32GB/s per chip
  – Device interface specified by JEDEC
High Bandwidth Memory (HBM)

- **Stacked Memory on SoC Architecture**
  - 4 to 8 die stacked on an SoC device
  - TSVs are typically employed to stack the memories
  - HBM stack then mounted on a 2.5D interposerer with a processing element – 1st key application is graphics
High Bandwidth Memory (HBM)

- **Micro Bump interface – defined by JEDEC**
  - A field of 3982 Micro Bumps
  - Micro bumps have a pitch of 27.5 x 48 staggered
  - Some micro bump locations are depopulated to permit test pads

![Diagram of HBM DRAM Die, Logic Die, and PHY with micro bumps and dimensions](image-url)
HBM Stack Probing

- Bottom of SoC device in the stack provides test pads in the field of Micro bumps
Probing challenges

- **Challenges:**
  - Probe without damaging Micro Bumps
  - No issue with FormFactor MicroSpring®
UltraFLEX KGS Solution

• **Test cell configuration:**
  – 36-slot UltraFLEX platform
  – 24 HPM 2.8Gbps digital instruments (3,480 IO pins)
  – 10 UVS256 device power instruments (2,560 VS pins)
  – New high-performance direct cable probe interface – no PIB or probe tower in the signal delivery path

• **Probe card based on proven Magnum 2x “P52”**
  – Mechanical standard already familiar to Form Factor
  – Digital and power pin assignments unique to UltraFLEX ATE
UltraFLEX KGS Enabling Capabilities

• **Digital**
  – Instrument delivers very fast signal rise times (<60ps)
  – Peaking options to compensate skin losses in path
  – In combination with probe card, full 2.8Gbps data rate at the die can be achieved

• **Device Power**
  – Programmable bandwidth to optimize response time of the supply and stability
  – Solves excessive droop issue seen on other ATE
Power Distribution Challenges

• Very high current on core supply
  – >4A per stack
  – Single power plane in the SoC device

• Multiple power levels required

• Power probe count may be limited
  – ~30 in the subject design

• Up to X96 parallelism
  – Large number of independent DUT power supplies
  – Power net routing on the PCBA and in the DUTlet
PDN Spring Layout

- Single VDD1 power plane on the DUT
- 3 power planes from the tester
  - Each ATE power plane is 2 DUT power supplies ganged to act as a single supply
  - Each ganged supply set has one sense connected at the DUT
- To equalize current from each ATE supply springs are interleaved
  - The 3 ganged power sources combine for a total of 4.2A on the DUT
PDN DUTlet Power Stack-up

- The sequence of 3 springs Power plane 1, 2 and 3, rotates around the DUT
- DC Current carrying per spring with no spring damage at 85°C is 500mA (1.4A capability per ATE power plane, 4.2A per DUT)

LGAs to PCBA Layer

Power plane 1
Power plane 2
GND
High Speed Signal routing
GND
Power Plane 3
GND

Multiple vias from each power plane to separate LGA – one via LGA for sense for each plane

Spring Attach Layer

3 adjacent springs each going to a different Power plane (total of 9 groups of 3 springs)

June 7-10, 2015
SW Test Workshop
PDN Performance Results

- Two approaches were used to evaluate PDN performance
  - Physical design data was extracted into an S-Parameter files using Cadence Sigrity SI tool
  - Actual measurements were done with a dual port VNA at the DUTlet on the full probe card
  - Results were simulated in the frequency domain using Agilent ADS simulator

- To address the concern about power sharing due to imbalance between the ATE power planes a 1 mOhm resistance was simulated to evaluate this impact
PDN simulations vs. Measurements

- Difference between extraction and VNA measurements is related to VNA measurements were done on decoupling capacitors as probe points
- Performance in the 10MHz to 100Mhz range correlates reasonably well
**PDN time domain simulation**

- Dual Port VNA measurements Port 1 on Tester LGA and Port 2 on decoupling cap
- VNA S parameter files were modified to include DC-component with 0.1 ohm of path resistance
PDN voltage drop across DUT

- Dual Port VNA measurements Port 1 on Tester LGA and Port 2 on decoupling cap
- VNA S parameter files were modified to include DC-component with 0.1 ohm of path resistance
Known Good Stack At Speed Test

- Functional test rate is nominally 500MHz / 1Gb/s
- Test modes and/or margin testing may demand higher test rates
- Test cell
  - FormFactor SM100 HFTAP with 1GHz capability
  - Teradyne UltraFLEX KGD with high speed memory instruments
  - Up to 96 sites can be tested in parallel at speed
Known Good Stack At Speed Test

- FormFactor’s Smart Matrix 100 HFTAP K10 and Teradyne's UltraFLEX KGD technology together deliver full Data Rate testing capability at wafer probe for Known Good Stack.
Summary

- FormFactor’s Smart Matrix 100 HFTAP K10 and Teradyne's UltraFLEX KGD technology together deliver full testing capability at wafer probe for Known Good Stack
- Power distribution concept has proven to be effective and does not show probe damage due to current imbalance
- Full at speed test has also been proven