IEICE TRANS. ELECTRON., VOL.E98–C, NO.4 APRIL 2015

PAPER Special Section on Solid-State Circuit Design—Architecture, Circuit, Device and Design Methodology

Relay Transmission Thruchip Interface with Low-Skew 3D Clock Distribution Network

Yasuhiro TAKE[†a)], ***Student Member and*** **Tadahiro KURODA**[†], ***Fellow***

SUMMARY This paper presents an inductive coupling interface using a relay transmission scheme and a low-skew 3D clock distribution network synchronized with an external reference clock source for 3D chip stacking. A relayed transmission scheme using one coil is proposed to reduce the number of coils in a data link. Coupled resonation is utilized for clock and data recovery (CDR) for the first time in the world, resulting in the elimination of a source-synchronous clock link. As a result, the total number of coils required is reduced to one-fifth of the conventional number required, yielding a significant improvement in data rate, layout area, and energy consumption. A low-skew 3D clock distribution network utilizes vertically coupled LC oscillators and horizontally coupled ring oscillators. The proposed frequency-locking and phase-pulling scheme widens the lock range to ±10%. Two test chips were designed and fabricated in 0.18 μm CMOS. The bandwidth of the proposed interface using relay transmission ThruChip Interface (TCI) is 2.7 Gb/s/mm^2; energy consumption per chip is 0.9 pJ/b/chip. Clock skew is less than 18- and 25- ps under a 1.8- and 0.9-V supply. The distributed RMS jitter is smaller than 1.72 ps.

key words: *TCI, Coupled-resonator, 3-D Integration, 3-D clock distribution, CDR*

1. Introduction

Three-dimensional (3D) integration is a key technology for creating LSIs with enhanced performance regardless of scaling. Up until recently, the performance of CMOS technology has been improved by scaling in accordance with Moore's Law. Owing to the physical limitations of scaling, however, the "end of Moore's law" is being discussed [1]. Thus, one way to improve CMOS performance regardless of scaling is drawing attention: 3D integration. By stacking chips in their lengthwise direction, systems can be expanded without suffering an area penalty. In addition, applying this stacking method enables the distance between the stacked chips to be reduced to the order of several dozen microns, thereby shortening the transmission distance between the chips. As the transmission distance is shortened in this manner, a high-speed and low-power-consumption interface between the chips—which is an essential component of high-performance LSIs—is created.

Applying a "ThruChip interface" (TCI) using inductive coupling makes it possible to integrate chips three-dimensionally by applying conventional CMOS technology without new additional processing [2], [3]. A TCI enables communication between chips by inductively coupling mutual on-chip coils. The coils are formed by metal wires and vias in a way that enables other metal wires to cross them. The transceiver layout area is equivalent to 36x that of a 2NAND gate. Although the additional cost of a TCI is much lower than that of a through-silicon via (TSV), speed, power, reliability, and testability are not compromised. Heterogeneous stacking (such as memory and logic stacking) can be done efficiently by AC coupling. Power can be delivered by conventional wirebonding. The development of the TCI has followed conventional ways of IC development in terms of design methodology, packaging, testing, and business scheme. In a similar manner to the digital CMOS performance improved by device scaling, LSIs will have their cost further reduced and their performance further improved by applying a TCI.

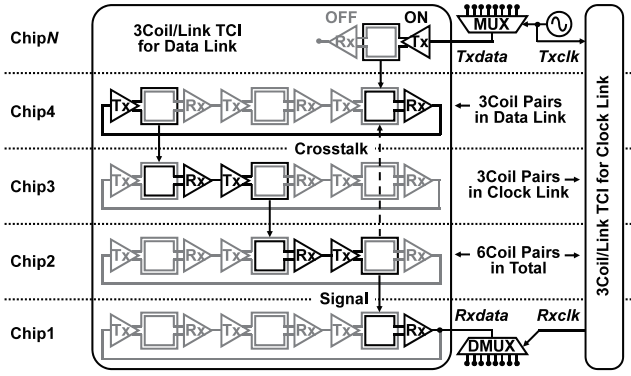


Fig. 1 Conventional ThruChip Interface.

The challenge that lies ahead for TCIs (namely, targeting 3D-integration systems with an operation clock rate exceeding the GHz range) is to expand bandwidth and accomplish low-skew clock distribution while reducing power consumption. To best use the performance of each chip, the bandwidth needs to be sufficiently broadened for communication between the chips. The bandwidth is expanded by parallelization of channels. However, the number of coils that can be placed in a certain area is restricted. As for the bandwidth of 2 Gb/s reported at ISSCC2010, three coils per data link are necessary to prevent interference between channels [4]. In addition, a source-synchronous clock link has three coils per channel, so six coils per channel are necessary as shown in Fig. 1.

Moreover, for sharing a common timing among all 3D-integrated chips, a low-skew clock distribution is required. When the operation clock rate exceeds the GHz range, a

Manuscript received July 21, 2014.
Manuscript revised November 30, 2014.
[†]The authors are with the Department of Electronics and Electrical Engineering, Keio University, Japan.
a) E-mail: take@kuroda.elec.keio.ac.jp
DOI: 10.1587/transele.E98.C.322

clock-timing margin of merely a dozen or so picoseconds is required. As for a conventional TCI, a source-synchronous clock link for synchronizing clocks is necessary. Increasing channel number to expand bandwidth increases the need for such a clock link as well as power consumption. More specifically, a synchronizer and a FIFO are needed, resulting in costs of significantly increased latency and degraded performance [5]. Thus, a method must be devised for distributing clocks while achieving low skew without increasing power consumption.

This paper reports the development of two test chips with a bandwidth five times that of the conventional TCI and a clock distribution scheme that can distribute a clock with phase and frequency synchronized with an external-reference clock to all chips in a 3D-integrated stack. This paper is an extended version of two previous works [6], [7], which complements the key technologies of the 3D-integration systems with an operation clock rate exceeding the GHz range, i.e., the model of the relay transmission TCI and coupled resonator. In the first test chip, the bandwidth was expanded by relay transmission TCI and 3D clock distribution using a LC coupled resonator. The relay transmission TCI reduces the number of coils of a data link from three to one. Moreover, the developed coupled LC resonator enables the three source-synchronous clock coils that had been necessary for each data channel to be reduced to one in the chip. As a result, the total number of coils required is reduced to one-fifth of the conventional number required. In addition, a transmit/receive timing model and an equation model of multiple coupled LC resonators were newly applied to the circuit architecture reported in [6]. Applying these models enables the process, transmission distance, and number of stacked chips to be selected flexibly. In the second test chip, a 1.1-GHz clock is distributed across stacked chips for the first time. It utilizes vertically coupled LC oscillators and horizontally coupled ring oscillators. The proposed frequency-locking and phase-pulling (FL-PP) scheme widens the lock range to ±10%, so the clock with phase and frequency synchronized with an external-reference clock can be delivered to all chips. Clock skew is less than 18- and 25- ps under a 1.8- and 0.9- V supply, and the distributed RMS jitter is smaller than 1.72 ps. Following up on a previous report [7], the usefulness of FL-PP was shown by comparing it with the prior art of clock distribution through wire bonding, and by analyzing lock ranges without FL-PP. In addition, further scalability was achieved by combining the FL-PP scheme with the relay transmission scheme in the first test chip.

The rest of the paper is organized as follows. Section II describes the first test chip for expanding the bandwidth of the TCI and its optimal design. It is shown that optimum timing design can be achieved by applying the transmit/receive timing model and coupled LC resonator model. In Section III explains the second test chip of 3D clock distribution using a coupled resonator with a FL-PP scheme. The measurement results of two test chips are shown in Section IV. Finally, Section V presents the conclusion.

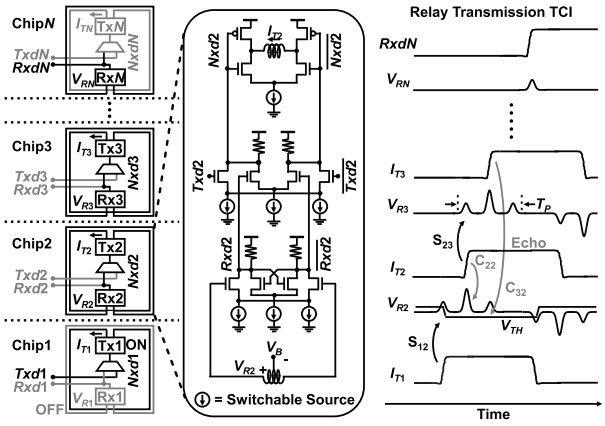


Fig. 2 Circuit diagram of relay transmission ThruChip Interface (TCI).

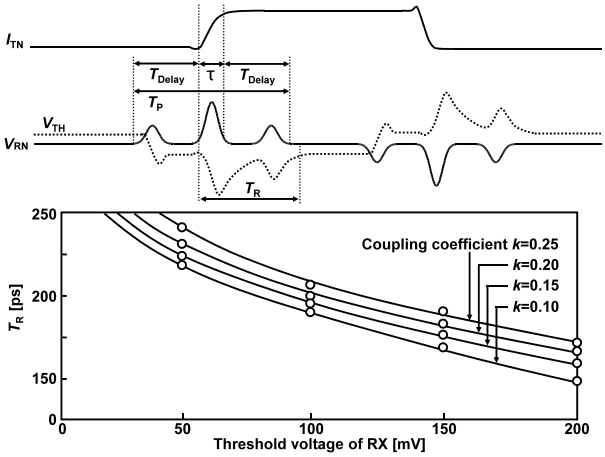


Fig. 3 (a) The transient response of receiver coil (b) Simulated recovery time of threshold voltage dependence on coupling coefficient.

2. Wide-Bandwidth TCI

The bandwidth is expanded by parallelization of channels. One method to expand the band is to reduce the number of the necessary coils per channel. In this paper, the bandwidth was increased up to five times by reducing the number of the necessary coils. Three data coils were reduced to one coil by a relay transmission TCI scheme. Three clock coils were deleted by CDR using the coupled LC resonator. Following up on the circuit concept reported in [6], the transmit/receive timing model (Section 2.1.2) and an equation model of the LC resonator (Section 2.2.2 and 2.2.3) were newly added.

2.1 Relay Transmission TCI

The relay transmission TCI reduces the number of coils of a data link from three to one while increasing the bandwidth by three times. A circuit diagram of a data link based on the relay transmission TCI is shown in Fig. 2. Switchable current sources that switch the function of the interface (transmitter, receive, or repeater) are newly added to the circuit in the conventional TCI transceiver shown in Fig. 1.

2.1.1 Circuit architecture

Figure 2 illustrates the case in which a controller is reading data from Chip 1 with assistance from the other chips as repeaters. Grey lines indicate circuits that are disabled momentarily by a program. When Chip 1 transmits signal S_{12}, a small pulsed voltage (V_{R2}) is induced and detected by Rx2 in Chip 2. The detected signal is amplified by the CML buffer and transmitted to Chip 3 by Tx2 as signal S_{23}, which induces an undesired second pulse in V_{R2}. That pulse is caused by self-crosstalk (C_{22}) and is larger than the first signal pulse due to shorter distance between Rx2 and Tx2, but it does not cause a malfunction. The Rx2 is implemented by a hysteresis comparator whose input threshold has been changed to receive the next signal with the opposite polarity. The Rx2 receives the third pulse crosstalk (C_{32}) when chip 3 sends the data to Chip 4. After that, Rx2 is ready to receive the next signal. Repeating this procedure enables signals to be transmitted in a relay manner on an arbitrary number of channels with one coil per channel.

2.1.2 Design of timing margin

To transmit signals correctly, the interval between the signals must be bigger than the cycle time (T_P) including crosstalk. T_P is an important parameter that determines the data rate of the relay transmission TCI. In this section, an equation model of relay transmission TCI is given to design the timing margin and maximum data rate. As shown in Fig. 3, T_P is given as the sum of delay time (T_{DELAY}) of the repeat circuit and received pulse width (τ) as follows:

$$T_P = 2T_{DELAY} + \tau \tag{1}$$

T_{DELAY} is in the order 100 ps in the case of the 0.18 μm process. Received pulse width (τ) is determined by dIT/dt derived from transmission current I_T. By rapid switching on the transmission side, τ can be shortened. In consideration of timing margin of the receiver and ringing noise due to channel gain, however, correct signal reception might not be possible. The frequency spectrum of the received pulse is a Gaussian distribution given as

$$|V_R(\omega)| = \frac{\sqrt{\pi}\tau V_p}{2} exp\left(-\frac{\omega^2\tau^2}{16}\right) \tag{2}$$

where V_P is amplitude of the received pulse. According to Equation (2), when τ becomes shorter, the frequency spectrum becomes wider. To transmit signals without distortion, the frequency band of the channels must be widened. Signal distortion is caused by not only communication error but also reduction of data rate due to increased T_P. The frequency range is given by the self-resonant frequency of the coil (f_{SR}), expressed as

$$f_{SR} = \frac{1}{2\pi\sqrt{LC}} \tag{3}$$

When the bandwidth for transmitting a signal is set, a value of f_{SR} exceeding that bandwidth is requested, and consequently the upper limit of coil inductance is determined. Mutual inductance M, obtained from coupling coefficient k and inductances L_{TX} and L_{RX} of the transmitting and receiving coils, respectively, given by Equation (4), is also given an upper limit.

$$M = k\sqrt{L_{TX}L_{RX}} \tag{4}$$

Here, k is a value uniquely determined by coil diameter and transmission distance. There is a lower limit of V_P, namely, the lowest value at which the receiver can correctly receive a signal. This lower limit is set in consideration of not only the sensitivity of the receiver but also error due to noise. V_P is expressed on the basis of pulse wide (τ) as

$$V_P = \frac{4}{\sqrt{\pi}} M \frac{I_P}{\tau} \tag{5}$$

Since a tradeoff exists between τ and M when the signal band is set, τ is also set. From Equations (1) and (5), Equation (6) is obtained as

$$T_P = 2T_{DELAY} + \frac{4}{\sqrt{\pi}} M \frac{I_P}{V_P} \tag{6}$$

The maximum bandwidth at which signal transmission is possible at one coil per channel (f_{max}) is given by Equation (7), namely, the inverse of T_P (given by Equation (6)), as follows:

$$f_{max} = \frac{1}{2T_{DELAY} + \frac{4}{\sqrt{\pi}} M \frac{I_P}{V_P}} \tag{7}$$

To attain that bandwidth, special attention must be paid to variation in the sensitivity of the receiver circuit. The large self-crosstalk with the same polarity shifts the threshold voltages in the receiver and degrades its sensitivity. Figure 3 shows a plot of the coupling coefficient k versus recovery time T_R until the threshold value returns. To receive the next incoming signal, the maximum bandwidth must be designed so that the following equation is satisfied.

$$f_{max} = \frac{1}{T_{DELAY} + T_R} \tag{8}$$

In this work, circuits are designed to operate with low-swing signal and achieve 0.4 ns of T_P, which corresponds to a data rate of 2.5 Gb/s. In this way, transceivers can be placed concentrically in the chip stacking. The number of coils to form a data link is reduced to one from the three that were needed to cope with the crosstalk problem in the conventional interface [4].

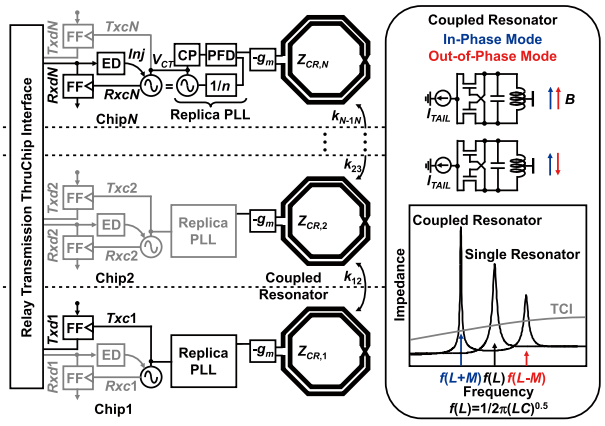


Fig. 4 Proposed CDR circuit using coupled-resonator.

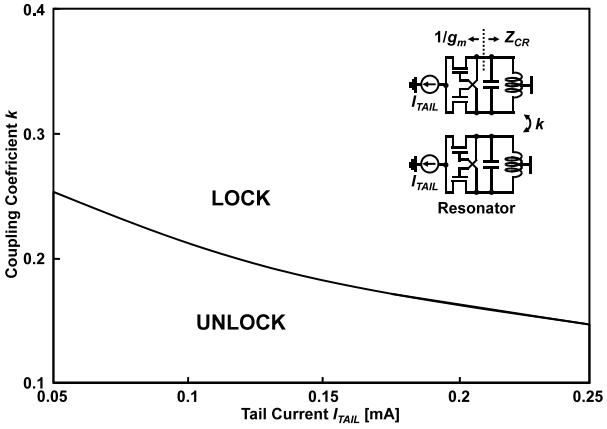


Fig. 5 Simulated locking range of coupled resonator dependence on tail current.

2.2 CDR using Coupled-resonator

2.2.1 Injection-locked CDR

The number of coils for data link is reduced by the relay transmission scheme. In addition, deleting the source-synchronous clock link effectively expands the bandwidth. Clock link can be eliminated by using clock and data recovery (CDR). The edge of the clock is synchronized to the data by injecting the edge signal of the data to a free-running oscillator. However, the CDR scheme requires global reference clock distribution among all the stacked chips. In addition, since this distribution network must always be active for random access, the power consumption needs to be reduced. In this investigation, coupled resonation [8], [9] is utilized for the reference clock link.

2.2.2 The model of single resonator

Figure 4 illustrates a circuit diagram of the clock link based on coupled resonation. LC oscillators generate synchronized clocks in each chip. All coils of LC oscillators are arranged to be piled up vertically. The clock outputs of the

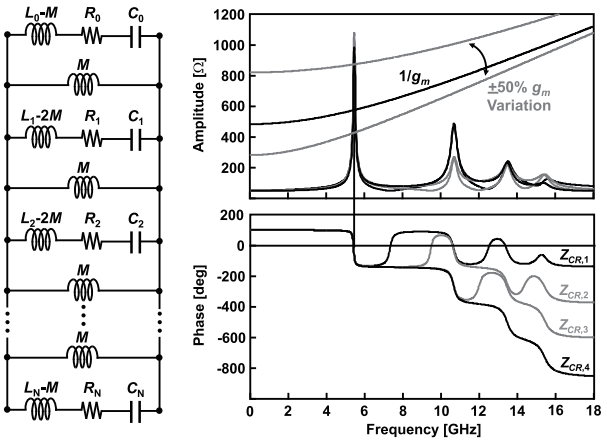


Fig. 6 Multiple stacked resonators (a) Equivalent model (b) Impedance characteristics (c) Phase characteristics.

coupled resonators are used as a reference frequency of a replica PLL in a receiver. The reference frequency is copied to a VCO by the PLL, and its phase is locked by injection at edges of received data [10]. The transimpedance between coils is changed by inductive coupling. When two coils are coupled, there are two peak points in characteristics of transimpedance as shown in Fig. 4. Transimpedance splits in the In-phase and Out-of-phase modes. When the values of capacitance and inductance in first and second resonators are the same, two resonant frequencies ω_A and ω_B are expressed as follows.

$$\omega_A = \frac{\omega_0}{\sqrt{1+k}}, \omega_B = \frac{\omega_0}{\sqrt{1-k}} \tag{9}$$

All resonators are parallel-connected, so the quality factor Q is given by

$$Q_A = \sqrt{1+k}Q_0, Q_B = \sqrt{1-k}Q_0 \tag{10}$$

The quality factor Q is $\sqrt{1+k}$ times than that of a single resonator Q_0. Therefore, the frequency and phase of a coupled resonator are both synchronized by inductive coupling between the oscillating coils with high Q factor. Since the coupling efficiency is very high due to the high Q factor, the power dissipation can be significantly reduced. Figure 5 demonstrates that the proposed resonator has sufficient design margin for coupling coefficient and tail current.

2.2.3 The model of multiple stacking resonator

The model of a multiple stacking resonator can be thought of as an equivalent circuit as shown in Fig. 6(a). Whenever the number of stacked chips increases, a series resonance circuit is connected in parallel. In the resonant situation, the sum of reactance equals zero. Therefore, the expression is as follows.

$$\frac{1}{\omega(L_0 - M) - \frac{1}{\omega C_0}} + \frac{1}{\omega M} + \frac{1}{\omega(L_1 - 2M) - \frac{1}{\omega C_1}}$$

$$+\frac{1}{\omega M}+\frac{1}{\omega\left(L_2-2M\right)-\frac{1}{\omega C_2}}+\frac{1}{\omega M}+$$
$$\cdots\cdot+\frac{1}{\omega M}+\frac{1}{\omega\left(L_N-M\right)-\frac{1}{\omega C_N}}=0 \quad (11)$$

where M is mutual inductance. When all values of capacitance, inductance, and mutual inductance of resonators are the same, the resonant frequency is as follows.

$$\frac{N-1}{\omega M}+\frac{2}{\omega\left(L-M\right)-\frac{1}{\omega C}}+$$
$$\frac{N-2}{\omega\left(L-2M\right)-\frac{1}{\omega C}}=0 \quad (12)$$

where N is the number of coupled resonators($N \geq 2$), $L = L_1 = L_2 = \cdots = L_N, C = C_1 = C_2 = \cdots = C_N$ The graph in the Fig. 6(b) shows simulated impedance (Z_{CR}) and phase characteristic of the coupled resonator when four LC oscillators are stacked. Four peaks appear in the impedance characteristic. In the first peak, the LC oscillators are coupled in the same phase. Since the oscillation condition of a crossed-coupled LC oscillator is given by $g_m Z_{CR} > 1$, Z_{CR} must exceed $1/g_m$ in order to oscillate the LC oscillators. The g_m can be designed so that the first peak only meets the oscillation condition even under large process variations. The coupled resonator oscillates just at the fundamental frequency and doesn't move to the other frequencies.

3. Low-skew 3D Clock Distribution Network

The clock timing between TX and RX can be synchronized by using CDR with a coupled resonator. However, there is some skew between each channel and circuit block because of the PVT variation and difference in the length of wire. This skew caused by mismatch limits the performance in 3D stacked chips. For sharing a common timing among all 3D-integrated chips, a low-skew clock distribution is required. When the operation clock rate exceeds the GHz range, a clock-timing margin of merely a dozen or so picoseconds is required. Making clock distribution with low-skew, low-jitter, and low-power is a significant challenge [11]. In this work, a 1.1-GHz clock is distributed across stacked chips for the first time. It utilizes vertically coupled LC oscillators and horizontally coupled ring oscillators. The proposed frequency-locking and phase-pulling (FL-PP) scheme widens the lock range to ±10%, so the clock with phase and frequency synchronized with an external-reference clock can be delivered to all chips.

3.1 Coupled Ring Oscillator for Horizontal Clock Distribution

In three-dimensional (3-D) integration, a low-skew, low-jitter, and low-power clock distribution is required in not only the inter-chip (vertical) but also the intra-chip (horizontal) clock distribution between stacked chips [12]. LC coupled resonators using inductive coupling through the oscillating coils deliver a clock to stacked chips with negligibly small skew. In contrast, conventional H-tree clock distributions are widely used for intra-chip clock distribution. However, the clock skew is increased by PVT variations associated with device scaling [13]. To replace H-tree distribution, a horizontal resonant clock distribution scheme is attracting growing interest. In particular, coupled ring oscillators that have shorted outputs [14] can reduce skew and jitter without additional layout area, unlike LC resonators [15]. The difference in phase and frequency within each oscillator caused by the PVT variations is equalized by mutual connection between the oscillators. Power dissipation can thus also be reduced since it is tolerant against variability due to low-voltage operation. A clock can be delivered to all stacked chips with negligibly small skew by combining LC coupled resonators for vertical distribution with coupled ring oscillators for horizontal distribution.

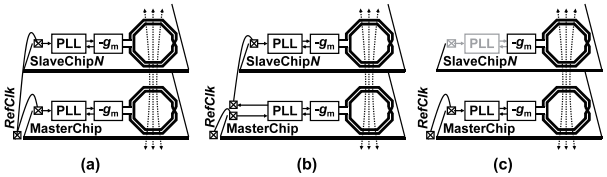


Fig. 7 Synchronization Methods (a) Control all varactor diodes (b) Provide control signal of the PLL (c) Control the varactor diode of Master chip.

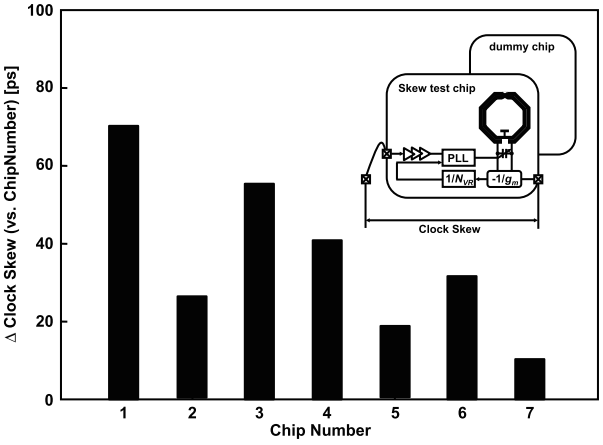
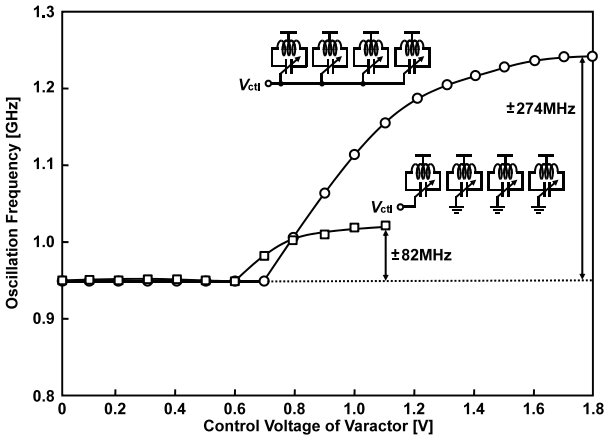


Fig. 8 Measured clock skew between bonding PAD and input of PLL.

3.2 Issue of Synchronization to Reference Clock

The resonance, however, is not synchronized with the reference clock. There are three possible synchronization methods, but they all have drawbacks.

- Control all varactor diodes of the LC oscillators with the PLL on each chip as shown in Fig. 7(a). This approach will yield large skew. For instance, suppose a reference clock is provided to a PLL on each chip by a bonding wire through buffer gates. A distributed clock will have some skew because of the delay variation in



s

Fig. 9 Simulated range of oscillation frequency dependence on control voltage of varactor.

the wire and the gates. Figure 8 indicates delay time of clock transmission from bonding pad on PCB to pad on chip. Seven test chips are fabricated by same design and process. The clock skew between bonding pad for reference clock on PCB and output clock synchronized with reference clock are measured. The measured clock skew is caused by variation in transmission path from bonding pad to input of PLL in chip. The maximum skew is measured to be 62 ps.

- Provide control signal of the PLL from a master chip to slave chips through bonding wires as shown in Fig. 7(b). This approach will yield large jitter since the bonding wires are very difficult to shield.
- Control the varactor diode of the LC oscillator on the master chip to pull in phase and frequency of the slave chips as shown in Fig. 7(c). The frequency range of a single varactor, however, is as small as ±4.3% since resonant frequency is given by total capacitance associated with the coupling as shown in Equation (12). Figure 9 shows that the frequency range of the coupled resonator depends on the number of controlled varactor diodes. Devices that have 10% variation will be difficult to produce. To mitigate these problems, a 3-D clock distribution is proposed that uses a Frequency-Locking and Phase-Pulling (FL-PP) synchronizer.

3.3 FL-PP Synchronizer

The FL-PP synchronizes the frequency and phase of whole stacked chips by a two- step control scheme combining Frequency-Looked Loop (FLL) with Phase-Looked Loop (PLL). The proposed 3D clock distribution using this scheme is illustrated in Fig. 10. LC-oscillators are inductively coupled through their coils for inter-chip (vertical) clock distribution, and ring oscillators are connected by wires at their outputs for intra-chip (horizontal) clock distribution. First (in Step 1), the frequency of the coupled LC oscillators is synchronized to the reference clock ($RefClk$) by a frequency locked loop (FLL) on each chip. In this step,

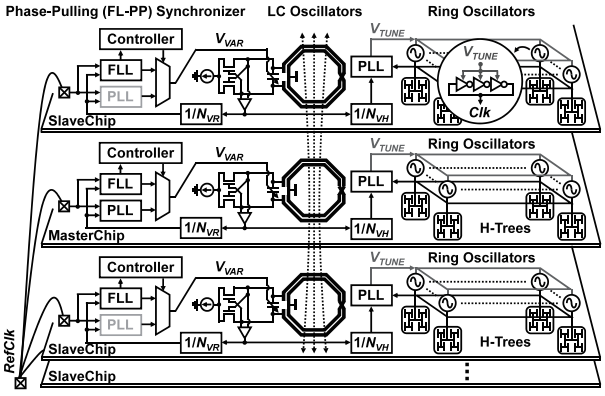


Fig. 10 3D clock distribution using Frequency-Locking and Phase-Pulling (FL-PP) scheme.

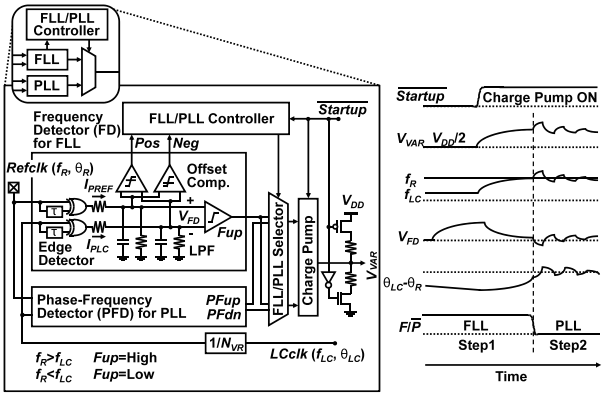


Fig. 11 FL-PP synchronizer circuit.

only the frequency is tuned, and the delay of the reference frequency of FLL due to the difference in wiring length is not a problem. As the varactor diodes of all the LC oscillators are changed, the variable range of frequency is widened. Second (in Step 2), the FLL is switched to a PLL in the master chip, and the phase is synchronized to the reference clock. The phases of the slave chips are also pulled in by inductive coupling. The loop gain of the FLLs in the slave chips is designed to be sufficiently weaker than that of the PLL in the master chip to prevent the FLLs obstructing the phase synchronization. The coupled ring oscillators can be used as a global clock distribution and connected to the H-tree that has gated clock buffers.

A circuit diagram and waveforms of the FL-PP synchronizer are depicted in Fig. 11. The reference clock and the output of the coupled LC-oscillators generate pulse-shaped currents, I_{PREF} and I_{PLC}, respectively, and their average difference is converted into a voltage V_{FD}. The frequency is controlled by a charge pump on the basis of the polarity of V_{FD}. When V_{FD} approaches zero, a FLL/PLL controller switches from Step 1 to Step 2. In many cases, the oscillation frequency of the LC oscillators is much higher than that of the ring oscillators when the layout area of the inductor is small. To balance the speed, the output of the LC oscillator needs to be divided by a $1/N$ frequency divider ($1/N_{VH}$ in Fig. 10). The $1/N$ frequency divider gen-

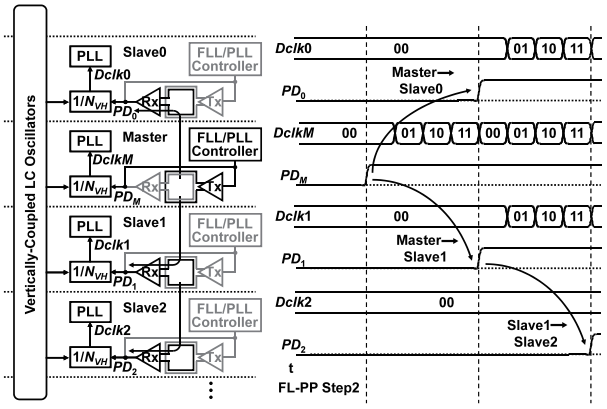


Fig. 12 Inductive-coupling clock divider synchronization by TCI.

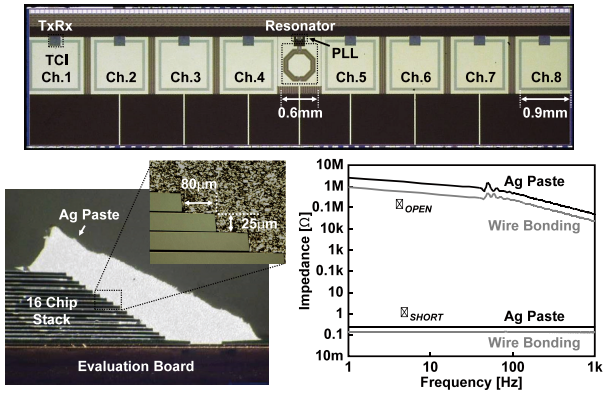


Fig. 13 Stacked test chips with Ag paste for power supply and measured supply Impedance.

erates a clock with an edge that changes at every N cycle of the input signal. The clock dividers need to be synchronized across stacked chips, which is very difficult because the input frequency is higher than 2 GHz.

3.4 Inductive-coupling Clock Divider

ThruChip Interface (TCI) [2], [3] is used for the synchronization as shown in Fig. 12. After completion of the phase pull-in in Step 2, a proceed-to-divide signal in the master chip (P_{DM}) becomes "High" to start the frequency divider. One input-clock before the next rising edge of the output, the P_{DM} is transferred to the upper and lower slave chips by TCI to generate P_{D0} and P_{D1}. With the same procedure, P_{DN} (N=0,1,. . .) will be relay transferred to all the slave chips. In this way, all the master and slave chips will be synchronized. Latency of the P_{DN} transfer is shorter than the cycle time of the LC oscillators (250 ps) as latency of TCI is less than 50 ps in a 0.18-μm CMOS.

4. Measurement Results

4.1 Relay Transmission TCI

A test chip was designed and fabricated in 0.18 μm CMOS (Fig. 13). Eight-bit parallel data channels are implemented

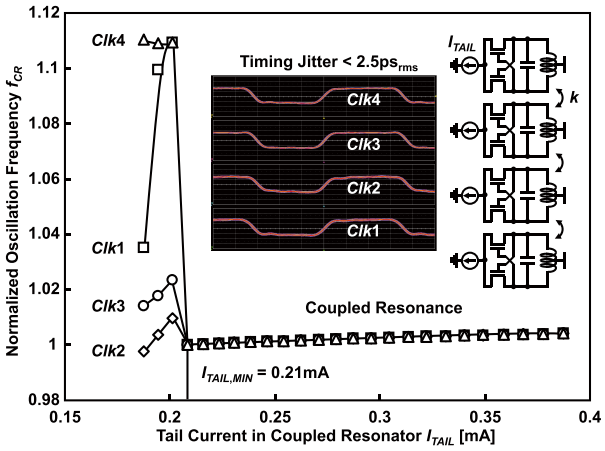


Fig. 14 Measured oscillation frequency dependence on tail current in coupled resonator.

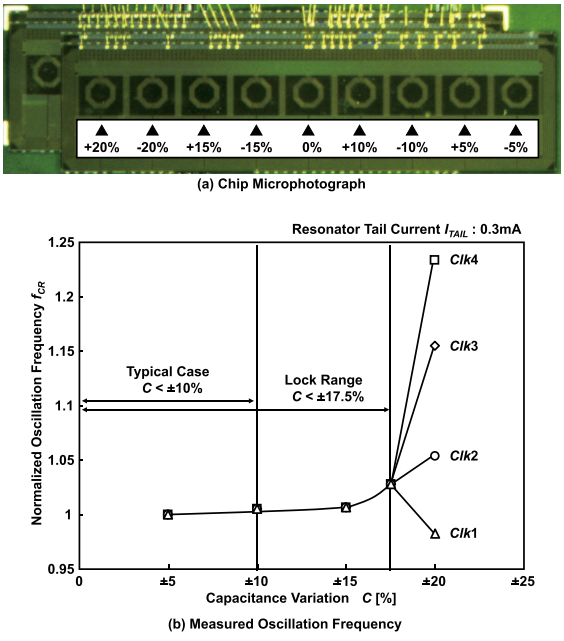


Fig. 15 Measured oscillation frequency dependence on capacitance variation.

with a coupled resonator and a replica PLL. Sixteen chips are stacked by terraced stacking with an offset of 80 μm. Silver (Ag) paste is used to supply power and ground. Cure temperature is 150°C , which is as low as that for conventional Die-Attachment Films (DAF). Processing time is only one minute. The sidewall of the chips is oxidized to prevent a short circuit between the Ag paste and the chip substrate. Previously, wire bonding was used where the chip offset needed be larger than 150 μm due to a large pad size and wire loop control. A large coil, typically 1.1 mm in diameter in this case, was therefore required to compensate for degradation in coupling caused by coil misalignment. The Ag paste is flexible enough to be applied to a 20-μm-wide small power pad. The chip offset was reduced by half, and hence, the coil diameter was reduced to 0.9 mm in the data link and 0.6 mm in the resonator. Measured

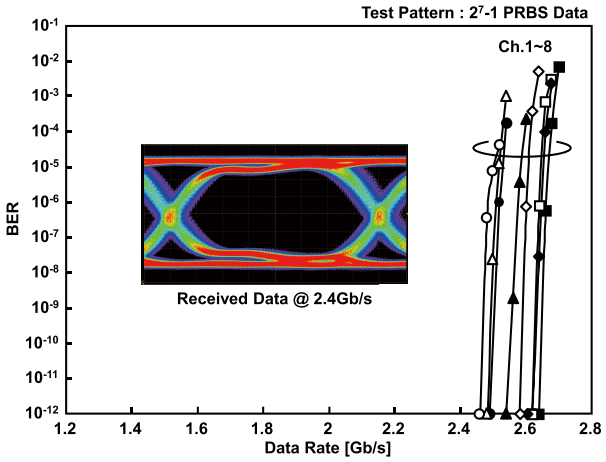


Fig. 16 Measured BER dependence on data rate.

impedance in Fig. 13 demonstrates that the Ag paste exhibits power supply integrity comparable to the wire bonding. Figure 14 shows measured oscillation frequency (f_{CR}) dependence on tail current (I_{TAIL}) in the coupled resonator. The coupled resonator successfully locks these four clocks though the largest discrepancy among clocks (in this case, between $Clk2$ and $Clk4$) is initially more than 10%. For the range of I_{TAIL} from 0.21 mA to 0.39 mA, four LC oscillators oscillate in the same frequency (f_{CR0}) and phase with a very small discrepancy of less than 0.01%. The measurement also demonstrates that the coupled resonator has up to ± 25% tolerance against V_{DD} change. Measured RMS timing jitter is less than 2.5 ps (< 2.4% U.I.), which is low enough to ensure BER lower than 10^{-12}. Power dissipation for clock generation and distribution is reduced to one-ninth that of the conventional scheme [4] where clock is generated at one chip and distributed to the other chips by relay transmission. To evaluate oscillation frequency dependence on LC variation, capacitors with different capacitance values (±20% of the standard value, in 5% increments) were implemented in the test chip as shown in Fig. 15(a). With two capacitors comprising one set, two sets were shifted in the X direction and stacked. By stacking an oscillator with +10% capacitance and one with −10% capacitance, it was confirmed that resonator mismatch was ±10%. As shown in Fig. 15(b), measured lock range on capacitance mismatch is ±17.5%, which is smaller than standard variation of Metal-Insulator-Metal (MIM) capacitors (±10%) in mass production. Frequency shift is less than 0.03% within the range of the ±10% capacitance variation. The rising slope in the graph is due to manufacturing mismatches in the test chips. Figure 16 shows measured BER dependence on the data rate. All eight channels in the test chip were tested by using $2^7 - 1$ PRBS data. BER<10^{-12} operation is confirmed at data rates higher than 2.4 Gb/s for all the channels. When there are long consecutive bits in the data pattern(e.g., 001111111111), it should be encoded in order to correct for any drift in the oscillator. The chip performance is summarized in Table 1. In conventional TCI six coils per channel have been necessary, but with the technology reported here

Table 1 Performance comparison of proposed relay transmission TCI.

		This Work	[4]Previous Work
Aggregated Bandwidth		19.2 Gb/s (9.6)	2.0 Gb/s (1)
Data Rate		2.4 Gb/s/channel	2.0 Gb/s/channel
Number of Data Coils		1 Coil/channel	3 Coils/channel
Number of Clock Coils		1 Coil/chip	3 Coils/channel
Coil Diameter		0.9 mm (Data) 0.6 mm (Clk)	1.1 mm (Data, Clk)
Total Layout Area		7.0 mm^2	7.3 mm^2
Bandwidth / Area		2.7 Gb/s/mm^2 (10)	0.27 Gb/s/mm^2 (1)
Energy Dissipation / Chip	Data	0.8 pJ/b	0.9 pJ/b
	Clock	0.1 pJ/b	0.9 pJ/b
	Total	0.9 pJ/b (1/2)	1.8 pJ/b (1)
Clock Recovery		CDR w/ Coupled Resonator	Source Synchronous

[4]M. Saito (ISSCC'10).

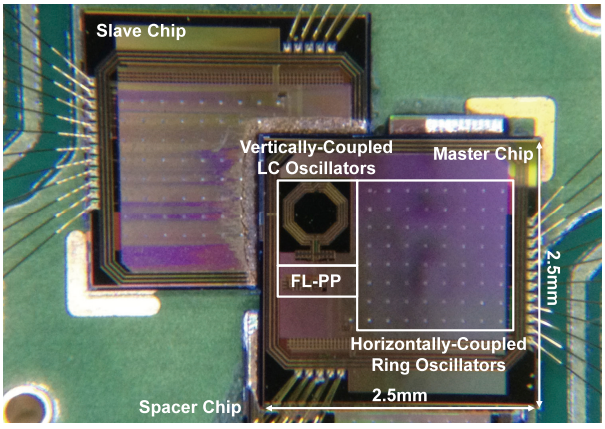


Fig. 17 Chip micrograph of stacked test chips.

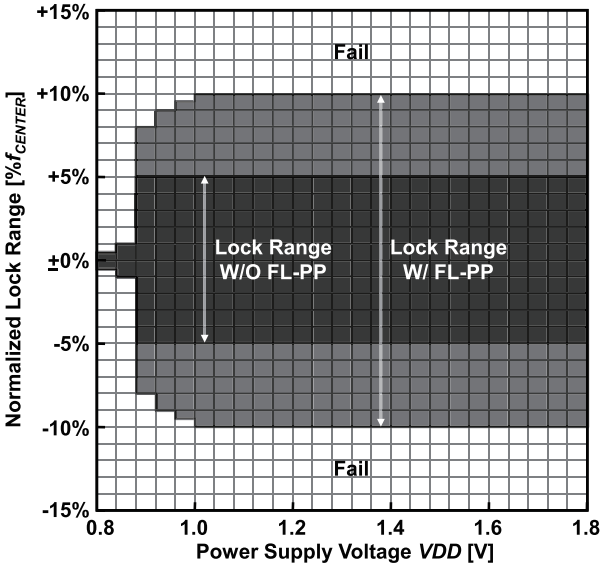


Fig. 18 Measured shmoo plot on lock range and power supply voltage.

the presence of six coils (5 coils for data links and 1 coil for clock link) enables the bandwidth to be expanded to five times that of the conventional TCI. In this test chip, the reduced channel diameter due to the use of Ag paste made it possible to increase the number of channels and expand the bandwidth 9.6 times.

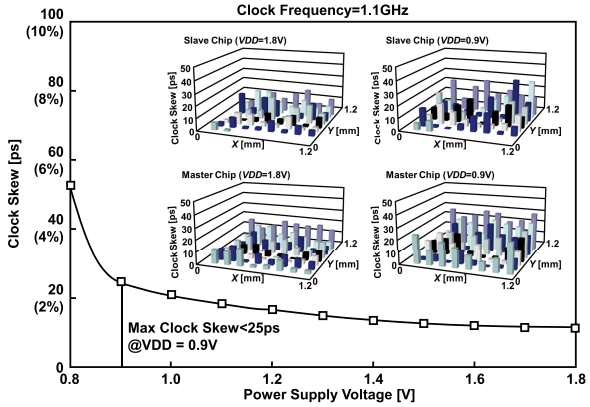


Fig. 19 Measured clock slew distribution and tolerance dependence on supply voltage.

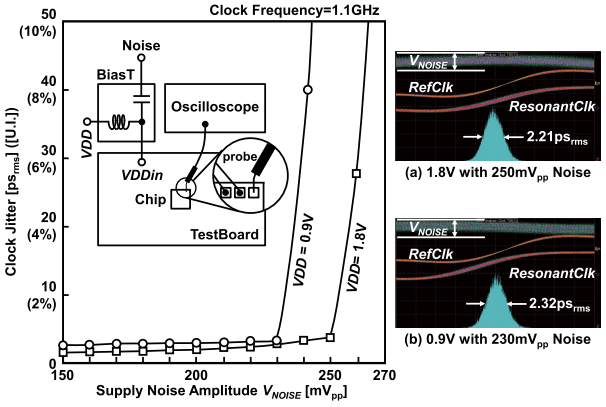


Fig. 20 Measured clock jitter dependence on power supply noise.

4.2 3D Clock Distribution using FL-PP

A test chip was designed and fabricated in a 0.18-μm CMOS process. A chip micrograph of stacked test chips is shown in Fig. 17. The diameter of the inductor is 600 μm for vertical clock distribution. A 40-μm thick master chip is stacked on a slave chip, both face-up and rotated, with 10-μm thick adhesive. Next, 64 ring oscillators are placed in a 175-μm pitch. Waveforms of each ring oscillator can be monitored by a picoprobe at 64 Pads through buffers. A measured shmoo plot of the lock range and power supply voltage is shown in Fig. 18. When there is FL-PP control, the lock range is extended to ±10%. It is almost two times wider than when there is not such control. In addition, when the supply power voltage V_{DD} for ring oscillators is lowered from 1.8 V to 0.9, the lock range is ±7.5%. Measured clock skew of whole stacked chips is shown in Fig. 19. In addition, Fig. 19 also shows the measured maximum clock skew dependence on supply voltage. The measured clock skew within all nodes of stacked coupled ring oscillators is less than 18 ps (1.8% U.I.) under the 1.8 V V_{DD} and 25 ps (2.5% U.I.) under the 0.9 V V_{DD}. The V_{DD} can be lowered to 0.9 V, at the cost of increasing skew to 7 ps. This skew is as low as the 2-D clock distribution technique for a microprocessor reported in [11]. 3-D clock distribution with the same skew as 2-D distribution can be realized between stacked chips. Measured RMS jitter shown in Fig. 20 is smaller than 1.63 ps_{rms} (< 0.2% U.I.) under the 1.8 V V_{DD} and 1.72 ps_{rms} (<0.2% U.I.) under the 0.9 V V_{DD}. To assess the impact of power supply noise on jitter, noise was intentionally applied to the test board power supply. An oscilloscope was used to observe the noise amplitude on the test chip power supply pad. Even when 260 mV white noise is intentionally added to the power supply, the measured jitter is less than 2.21 ps_{rms} (< 0.3% U.I.) under the 1.8 V V_{DD}. The jitter is only 2.32 ps_{rms} (< 0.3% U.I.) under the 0.9 V V_{DD} even when 230 mV white noise is added. Total power dissipation is 196 mW with the 0.9 V V_{DD}. Power dissipation per clock distribution area is 65.3mW/mm^2, which is 25% less than in a TSV and H-Tree distribution scheme [12]. In [12], three 3D clock distribution schemes using TSV a re compared. The block distribution area was 1 mm^2 smaller than that given here although the same 0.18-μm process was used. In the three schemes, this work was compared to the one combining H-tree topology with TSVs, which showed the lowest skew. Table 2 shows the performance comparison.

Table 2 Performance comparison of proposed clock distribution scheme.

	This Work	[10]	[9]
Clock Distribution Scheme	3D (Inductive+RO)	3D (TSV+H-Tree)	2D (H-Tree)
Clock Frequency	1.1 GHz	1.0 GHz	1.1 GHz
Clock Skew	<25 ps	<33 ps	<25 ps
Clock Jitter	<1.7 ps_{rms}	N/A	<5.0 ps
Power Dissipation	196 mW	260 mW	N/A
Clock Distribution Area	1.2 × 1.2 mm	1.0 × 1.0 mm	N/A
Power/Area	65.3 mW/mm^2	86.7 mW/mm^2	N/A
Process	180 nm CMOS	180 nm CMOS	180 nm CMOS

[9] Phillip J. Restle (ISSCC 2002).
[10] Vasilis F. Pavlidis (CICC 2008).

4.3 Impact of Combined Two Test Chips

A technique that combines the relay transmission scheme and the 3D clock distribution scheme using FL-PP makes designing a 3D integrated system whose operating frequency is over GHz easy; further, one does not need to be conscious of any severe timing limitation. Like a single chip, data communication is totally possible between 3D stacked chips, since the timing in whole stacked chips is synchronized. Figure 21 depicts the system that combines the relay transmission scheme and the 3D clock distribution scheme using FL-PP. The bandwidth can be significantly expanded while the distributed clock skew is kept under 18-ps if the communication distance is shortened. The graph in Fig. 22 shows the total bandwidth versus communication distance. In this simulation, the chip area is fixed to 1.2 mm × 1.2 mm since the clock skew of 18-ps in 1.2 mm × 1.2 mm chip area is guaranteed from measurement result. The bandwidth of 288 Gb/s with 18 ps low-skew clock distribution can be achieved when the communication distance is 20 μm. The communication distance less than 20 μm can be realized

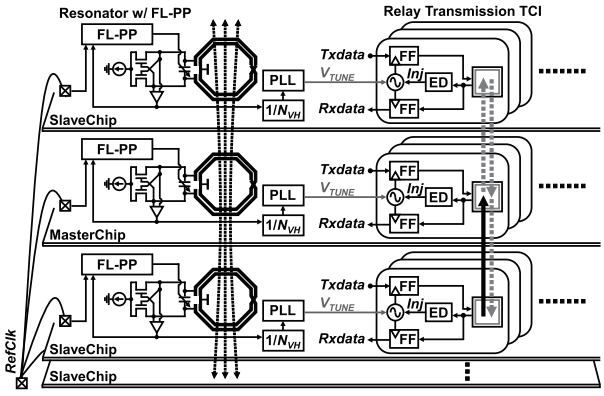


Fig. 21 Block diagram of the relay transmission scheme using 3D clock distribution with FL-PP.

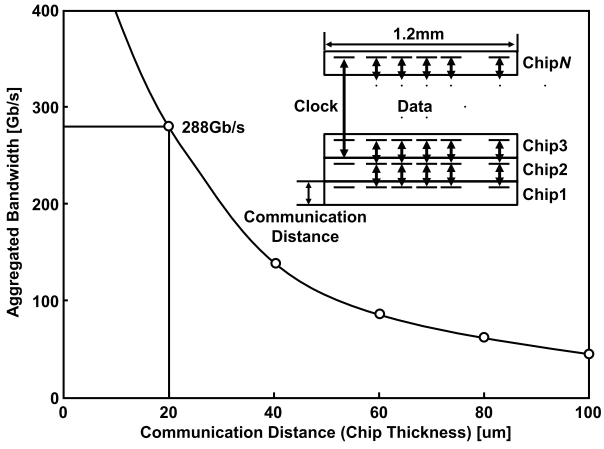


Fig. 22 Total bandwidth dependence on communication distance.

since the inductive coupling interface whose communication distance is less than 20 μm has been reported [16].

5. Conclusion

In this work, an inductive coupling interface using a relay transmission scheme and a low-skew 3D clock distribution network synchronized with an external reference clock source for 3D chip stacking were presented. The relay transmission TCI and CDR using a coupled resonator reduce the number of coils one-fifth of the conventional number required. As a result, the proposed interface achieves a bandwidth of 2.7 Gb/s/mm^2 and an energy consumption of 0.9 pJ/b/chip. A low-skew 3D clock distribution network delivers a clock synchronized with an external reference clock source to all of the stacked chips. Clock skew is less than 18- and 25- ps under a 1.8- and 0.9- V supply, and the distributed RMS jitter is smaller than 1.72 ps. The 288 Gb/s total bandwidth with 18 ps low-skew clock distribution can be achieved via a combined relay transmission scheme and a 3D clock distribution scheme using FL-PP.

Acknowledgments

This work is supported by CREST/JST.

References

[1] G. E. Moore, "No exponential is forever: But "Forever" can be delayed!," IEEE ISSCC Dig. Tech. Papers, pp.20–23, Feb. 2003.
[2] N. Miura, Y. Kohama, Y. Sugimori, H. Ishikuro, T. Sakurai, and T. Kuroda, "An 11 Gb/s inductive-coupling link with bust transmission," Proc. ISSCC, pp.298–299, Feb. 2008.
[3] N. Miura, Y. Kohama, Y. Sugimori, H. Ishikuro, T. Sakurai, and T. Kuroda, "A high-speed inductive-coupling link with burst transmission," IEEE J. Solid-State Circuits, vol.44, no.3, pp.947–954, Mar. 2009.
[4] M. Saito, N. Miura, and T. Kuroda, "A 2 Gb/s 1.8 pJ/b/chip inductive-coupling through-chip bus for 128-Die NAND-flash memory stacking," ISSCC Dig. Tech. Papers, pp.440–441, Feb. 2010.
[5] J. Kim, I. Verbauwhede, and M.-C. F. Chang, "Design of an interconnect architecture and signaling technology for parallelism in communication," IEEE Transactions on VLSI Systems, vol.15, no.8, pp.881–894, Aug. 2007.
[6] N. Miura, Y. Take, M. Saito, Y. Yoshida, and T. Kuroda, "A 2.7 Gb/s/mm^2 0.9 pJ/b/Chip 1 Coil/Channel ThruChip Interface with coupled-resonator-based CDR for NAND flash memory stacking," ISSCC Dig. Tech. Papers, pp.490–491, Feb. 2011.
[7] Y. Take, N. Miura, H. Ishikuro, and T. Kuroda, "3D clock distribution using vertically/horizontally coupled resonators," ISSCC Dig. Tech. Papers, pp.258–259, Feb. 2013.
[8] R. Adler, "A study of locking phenomena in oscillators," Proc. IEEE, vol.61, no.10, pp.1380–1385, Oct. 1973.
[9] A. Mazzanti and F. Svelto, "A 1.8-GHz injection-locked quadrature CMOS VCO with low phase noise and high phase accuracy," IEEE Trans. Circ. Syst., vol.53, no.3, pp.554–560, Mar. 2006.
[10] N. Miura, K. Kasuga, M. Saito, and T. Kuroda, "An 8 Tb/s 1 pJ/b 0.8 mm^2/Tb/s QDR inductive-coupling interface between 65 nm CMOS and 0.1 μm DRAM," ISSCC Dig. Tech. Papers, pp.436–437, Feb. 2010.
[11] P. J. Restle, C. A. Carter, J. P. Eckhardt, B. L. Krauter, B. D. McCredie, K. A. Jenkins, A. J. Weger, and A. V. Mule, "The clock distribution of the power 4 microprocessor," ISSCC Dig. Tech. Papers, pp.144–145, Feb. 2002.
[12] V. F. Pavlidis, I. Savidis, and E. G. Friedman, "Clock distribution networks for 3-D integrated circuits," CICC, pp.651–654, Sept. 2008.
[13] S. Nassif, "Delay variability: Sources, impacts and trends," ISSCC Dig. Tech. Papers, pp.368–369, Feb. 2000.
[14] H. Mizuno and K. Ishibashi, "A noise-immune GHz-clock distribution scheme using synchronous distributed oscillators," ISSCC Dig. Tech. Papers, pp.404–405, Feb. 1998.
[15] S. C. Chan, K. L. Shepard, and P. J. Restle, "Distributed differential oscillators for global clock networks," IEEE J. Solid-State Circuits, vol.41, no.9, pp.2083–2094, Sept. 2006.
[16] N. Miura, D. Mizoguchi, M. Inoue, T. Sakurai, and T. Kuroda, "A 195-Gb/s 1.2-W inductive inter-chip wireless superconnect with transmit power control scheme for 3-D-stacked system in a package," IEEE J. Solid-State Circuits, vol.41, no.1, pp.23–34, Jan. 2006.

Yasuhiro Take received the B.S. and M.S. degree in electrical engineering from Keio University, Yokohama, Japan, in 2010 and 2012 respectively where he is currently working toward the Ph.D. degree. Since 2009, he has been engaged in research on the 3-D-stacked inductive interchip wireless interface for system in a package.

Tadahiro Kuroda received the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999. In 1982, he joined Toshiba Corporation, where he designed CMOS SRAMs and ASICs. From 1988 to 1990, he was a Visiting Scholar with the University of California, Berkeley, where he conducted research in the field of VLSI CAD. In 1990, he was back to Toshiba, and engaged in the research and development of BiCMOS/ECL ASICs, high-speed CMOS LSIs for telecommunications and low-power CMOS LSIs for mobile applications. He invented a Variable Threshold-voltage CMOS (VTCMOS) technology to control VTH through substrate bias, and applied it to a DCT core processor in 1995. He also developed a Variable Supply-voltage scheme to control VDD by an embedded DC-DC converter, and employed it to a microprocessor core and an MPEG-4 chip in 1997. In 2000, he moved to Keio University, Yokohama, Japan, where he has been a professor since 2002. He was a Visiting MacKay Professor at the University of California, Berkeley in 2007. His research interests include low-power, high-speed CMOS design, proximity communications using inductive/EM-coupling and image recognition. He has published more than 200 technical publications, including 34 ISSCC papers, 21 VLSI Symposia papers, 19 CICC papers and 16 A-SSCC papers. He wrote 22 books/chapters and filed more than 100 patents. Dr. Kuroda served as the General Chairman for the Symposium on VLSI Circuits and A-SSCC, the Vice Chairman for ASP-DAC, sub-committee chairs for A-SSCC, ICCAD, SSDM and VLSI-DAT, and TPC members for ISSCC, the Symposium on VLSI Circuits, CICC, DAC, ASP-DAC, ISLPED, SSDM, ISQED, and other international conferences. He is a recipient of the 2005 P&I Patent of the Year Award, the 2007 ASP-DAC Best Design Award, the 2009 IEICE Achievement Award, and the 2011 IEICE Society Award. He was an elected AdCom member, a Distinguished Lecturer, and a representative of Region 10 for the IEEE Solid-State Circuits Society. He is an IEEE Fellow and an IEICE Fellow.