A Scalable and Reconfigurable 2.5D Integrated Multicore Processor on Silicon Interposer

Jie Lin[1], Shikai Zhu[1], Zhiyi Yu[1, 2], Dongjun Xu[3], Sai Manoj P.D.[3], Hao Yu[3]

[1] State Key Lab of ASIC and System, Fudan University, Shanghai 201203, P.R.China

[2] SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University, Guangzhou 510006, P.R.China

[3] School of EEE, Nanyang Technological University, 639798, Singapore

***Abstract* — This paper presents a novel 2.5D multicore processor which consists of 3 distinct silicon dies: a processor die with 8 MIPS-cores, a 16kB SRAM die, and an accelerator die for multimedia and communication applications. These dies are interconnected into multi-modes, like core-core (up to 32 cores), core-memory (4x storage capacity) and core-accelerator (4.4x speedup in H.264 decoder), to establish a scalable and reconfigurable platform with less tape-out die area cost. A pair of 8Gbps SerDes is custom designed for each of the 12 inter-die communication channels, achieving a 2.5D I/O bandwidth of 24GB/s. The processor was implemented in GF 65nm process, and operates at 500MHz under 1.2V supply, with 1.08W power dissipation.**

***Index Terms* — 2.5D stacking, multicore processor, TSI, through-silicon interposer, SerDes, high bandwidth.**

I. Introduction

With the technology scaling and design methodology improvement, number of processing cores has been increasing progressively on each die to scale performance while keeping power in check [1]. For homogeneous multi-core processor, it is usually implemented by layout copy for architectural expansion in an easy way, while lacking high performance in various situations. Yet, heterogeneous multicores are normally integrated with dedicated accelerators to outperform the former with a certain design complexity. This fact motivates us to explore a novel multicore architecture with the advantage of scalability and reconfigurablility at low implementation cost.

In [2], a 2.5D system integration scheme is proposed, which integrates several partitioned dies side by side on a common substrate. And it greatly suits for our exploring architecture: adopting a processor die with 8 homogeneous MIPS cores as principal part and an off-chip memory die for capacity expansion, together with an accelerator die, which helps in faster kernel processing in multimedia and communication applications. The design effort focuses on implementing a high bandwidth and low latency reconfigurable interface between dies.

A growing number of scientific research and industrial products involving 2.5D technology have been announced in recent years. Academic papers exploited the design space in the enriched concept of heterogeneous integration and performed system modelling and simulation before 2011, in which year Xilinx released its industry-first product, which bonded 4 homogeneous FPGA slices onto a silicon interposer, increasing FPGA volume vastly [3]. During these years, research works reported on 2.5D system with physical implementation trickled in regarding several fields. In [4], an early prototype of integrated voltage regulator using 2.5D stacking for power inductor integration was presented with effective implementation of DVFS. Ref. [5] introduced a 2.5D heterogeneously integrated bio-sensing microsystem, well applied in multi-channel neural-sensing.

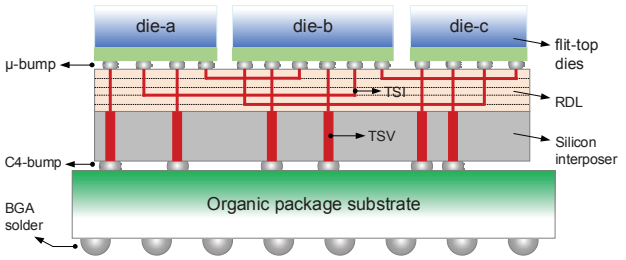

Fig. 1. 2.5D integrated system diagram

The 2.5D assembly of this work is implemented by silicon interposer with TSV (through-silicon via) and is to be manufactured by national center for advanced packaging (NCAP, China). As shown in Fig. 1, the flip-top dies are mounted on silicon interposer side by side with an array of micro-bumps (μ-bump). Some of the μ-bumps are interconnected by metal wire in RDLs (redistribution layers) for inter-die communication, which do not require large I/O pads, thus saving area, while others are linked to C4 bumps through TSVs, routing P/G and I/O signals for conventional BGA package. The way spread chips on makes 2.5D stacking outperform 3D vertically stacking mainly in terms of thermal dissipation.

II. System Architecture

The overview system architecture of the proposed 2.5D multicore processor is shown in Fig. 2. It basically has an

978-1-4799-8682-8/15/$31.00 ©2015 IEEE

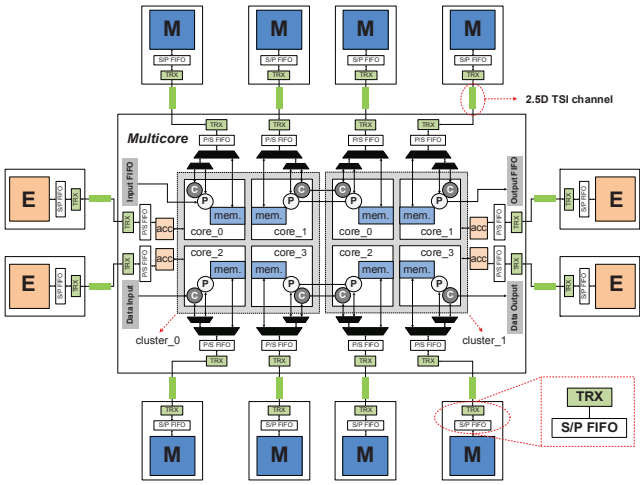

Fig. 2. Architecture of the proposed 2.5D multicore processor

8-core MIPS microprocessor, which can work as a minimum system. Inherited from our previous work in [6], the cores communicates through a high bandwidth and power efficient 2D mesh network-on-chip (NoC), which features packet-controlled circuit-switched double-layer routing. For the sake of data locality, the cores are further divided into 2 clusters, either of which has an individual clock domain with fine-grain clock-gating. Each MIPS core is designed with private instruction memory (6KB) and shared on-chip data memory with small capacity of 4KB.

To meet the requirement of storage capacity in data-intensive applications, the system is integrated with 8 identical off-chip memory blocks longitudinally, denoted by **M** in Fig. 2, with each having a capacity of 16KB SRAM and can be accessed by pipeline or direct memory access (DMA) engine.

In order to process kernels faster in multimedia and communication applications, this work further employs 4 identical expanded accelerator dies (denoted by **E**) in the transverse direction, each consisting of entropy decoder for H.264 and 16-point FFT, complex multiplier for LTE. Obviously, only accelerator die needs to be redesigned and manufactured in case of altering applications.

Core-memory and core-accelerator connections adopt the same interface circuitries, except that the former one employed additional multiplexer logic, making it support core-core connection by double-layer NoC expansion, thus obtaining multiple computational power and beneficial to the applications like big data, neuroscience computing, etc. The ports of packet/circuit switched routers located on left/right sides are assigned to system I/O, which limits the core-core expansion transversely in this prototype. The 2.5D through-silicon interposer (TSI) channel is indicated by green colored rectangular boxes in Fig. 2.

Apart from the reconfigrurability of the interface, the design targets on low latency and high bandwidth of the inter-die communication, which will be explained in the following sections.

III. Interface Circuitry Between Dies

The interface circuitry is a critical part in the 2.5D integrated system, and it is primarily dominated by the limitation of IO resources for inter-die connection, since all the I/O cells (providing driver and ESD for package IO) and TRX modules ought to be routed to the landing metal block on the top layer (as shown in Fig. 7) for further μbump-bonding. The octagonal metal block is drawn according to both the design rules of GF library and requirements of μbump technology, with height, width and pitch of 75 μm, 75 μm and 160μm respectively. This size constrain limits the total number of landing metal block to 246 on multicore die, within which the number is reduced to 126 for conventional package bonding just to allocate more I/O resources to inter-die connection. As a result, only 5 TSI I/Os are available for one-way transmission, when allowing 12 bidirectional inter-die communication in parallel.

A. Digital Circuits

The overall interface circuitry between dies is illustrated in Fig. 3, in which the digital circuits are composed of asynchronous FIFO, digital Parallel-Serial & Serial-Parallel circuit, and error code detection & correction module. At system boot phase, the configuration register of "*expansion mode*" is initialized, together with "*coordinate of router*". The 32-bits packetized data from double-layer NoC (at core-core expansion mode) or off-chip memory access (at core-memory expansion mode) is selected by a multiplexer, and then stored in an asynchronous FIFO for digital-analog clock domains isolation. Parallel-Serial (32bits-8bits) and Serial-Parallel (8bits-32bits) logic circuits are inserted on the basis of the I/O quantity limitation. The 8-bits data is further transmitted through a pair of TSI differential channel at the speed of 8Gbps, handled by an analog SerDes. Along with the serialized data, 3 bits of control signals are separately transmitted – *w_en, w_full, w_index*: the first two are for FIFO write protocol, and the last one is for particular notification decided by designer.

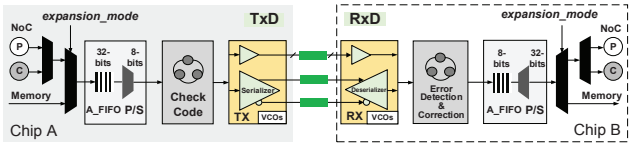

Fig. 3. Interface circuitry between dies

To solve the possible bits disorder problem at the output of *Deserializer*, a set of circuit is designed, which is always activated at the boot phase and then outputs data with corrected bits' order by sending and detecting a serial of check code.

B. Custom-Designed TRX

The custom-designed TRX consists of 3 buffers driving low-speed control signals, as well as a SerDes transferring 8-bits data, the circuit of which is depicted in Fig. 4.

Firstly, the 8-bits input data provided at the transmitter is serialized. The serializer is designed using the D flip-flop followed by a multiplexer. A current-mode logic (CML) buffer drives the TSI-based transmission line (T-line). A voltage-controlled oscillator (VCO) is deployed to maintain the clock frequency, at which data is transmitted.

A sampler at the receiver is utilized to sample the received signals and convert them into digital signals. This is followed by a delay-locked loop (DLL) based clock-data recovery (CDR) circuit to de-skew the sampling clocks. Two exclusive-or (XOR) gates in Fig. 4 (inset right bottom figure) form a phase detector to judge the sampling clock position compared to input data and provide "early" pulse and "late" pulse. A charge-pump block converts these pulses into a variable voltage to control the DLL delay line, which can tune the delay phase of clocks and also provide feedback to the sampler.

The post-layout simulations show that the transmitter/receiver consumes a power of 15.2mW/7.1mW with a delay of 1.16ns/2.67ns separately at the speed of 8Gbps, and the delay of 2.5D TSI channel is 49.3ps in T-line model for the length of 3mm.

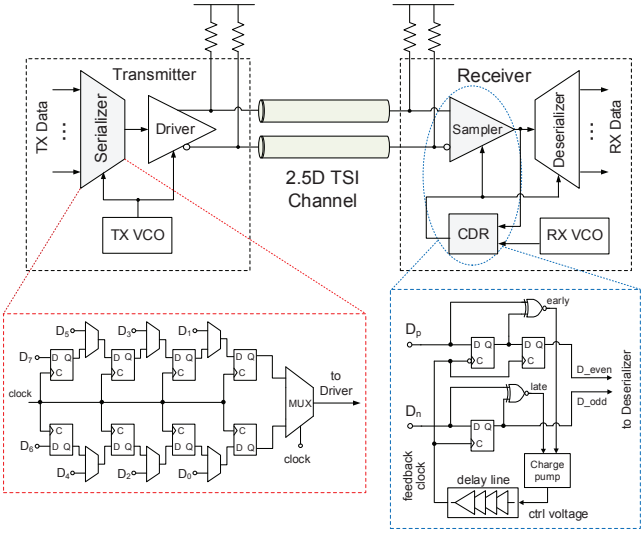

Fig. 4. The circuit of custom-designed SerDes

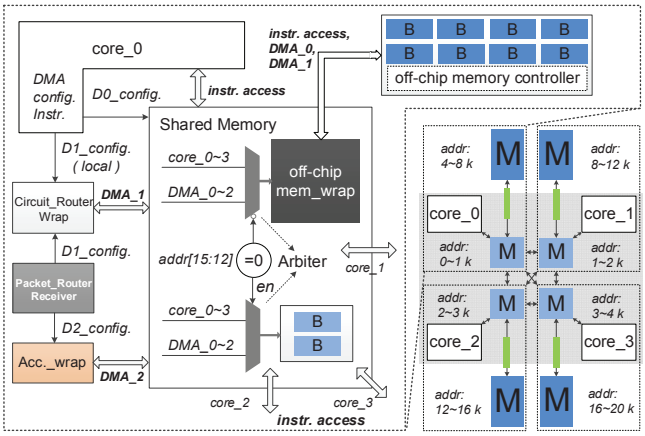

Fig. 5. Extended memory system

IV. Extended Memory System and Management

All on/off-chip data memory blocks are shared by cores within the same cluster, and DMA operation is supported to transfer streaming data background between any-to-any on-chip memory and off-chip memory/accelerator, including across chips in core-core expansion mode. As illustrated in Fig. 5, all access requests from cores and DMAs are served by an arbiter with round-robin algorithm to avoid deadlock. On/off-chip SRAMs are further split into 0.5k word/bank, *i.e.* 2k byte/bank, to improve memory access bandwidth and shorten the critical timing path. The address coding starts from on-chip memory and then off-chip clockwise, extending cluster-shared storage space from 4k to 20k word.

The on-chip memory is utilized to store run-time program variables and synchronization flags between cores, due to its low access latency (one clock cycle for local, and two for remote), while the off-chip memory is for streaming data storage in embedded applications, whose large interface latency can be significantly hidden by DMA. Fig. 6 illustrates the local access latency to the off-chip memory, and it shows that DMA does not occupy pipeline time since it transfers data background once the special configuration instructions are executed. Besides, DMA requires less packet head and thus consumes less data transfer time. The DMA write/read saves up to 49.8% and 90.3% transfer time (latency) respectively at 512 words compared with pure instruction write/read.

As the data characteristic of our specific application is mostly continuous and stream-like, the various DMA mechanisms act as hardware prefetching triggered by special instruction. This cache-free memory hierarchy could avoid coherence issues and maintain system simplicity [7].

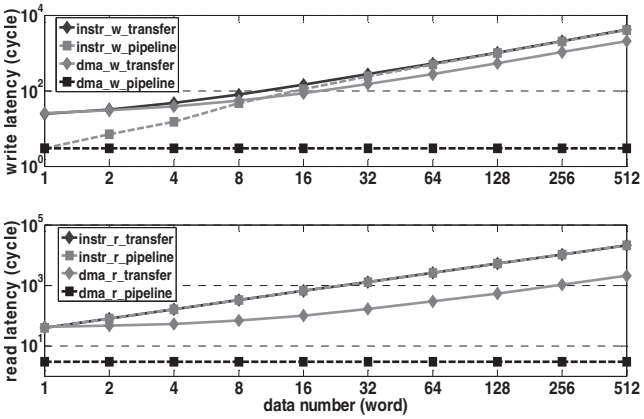

Fig. 6. Local access latency to off-chip memory

V. Results

Fig. 7 presents the photo of the three dies and their characteristics obtained from post-layout simulation, since the manufactured dies are pending on 2.5D assembly by NCAP. To demonstrate the performance of this 2.5D integrated multicore processor, an H.264 intra decoder is mapped to 8 cores. By calling entropy decoder in off-chip accelerator die, which features coarse-grained and loosely coupled with pipeline, the decoder performance achieves 720p@34fps at 500MHz, which is 4.4x better than pure software solution, and also 1.7x better per core on average than its predecessor [6].

Additionally, a typical big data application - clustering algorithm is implemented by 4 cores basing on a 16000 point three-dimensional array of dataset, utilizing canopy method proposed in [8]. Simulation results show that after well-inserted DMA operations in accessing off-chip memory, the overall computational time has been reduced 43.8% compared to instruction read/write. Besides, it achieves 3.8x performance improvement after mapped into 16 cores when expanding an extra 8-cores processor, which exhibits the flexibility and scalability of the proposed 2.5D integrated architecture.

VI. Conclusion and Acknowledgment

This paper presents a 2.5D integrated multicore processor, which consists of 3 distinct silicon dies. They are flexible to be organized into various multi-chip systems to meet different application requirements, thus saving none recurring engineering (NRE) costs and shortening time-to-market. The multicore chip supports 12 way full-duplex communication in parallel, bringing the bandwidth up to 24GB/s. Simulation results show that the specific applications have achieved better performance owing to hardware acceleration, cores scaling, as well as software-aided DMA which significantly hides the interface latency.

Technology	GF 65 nm LPE	**μ-bump**	246
Frequency	500MHz (1.2V)	**2.5D IO Speed**	8Gbps (max)
Typical Power	1.08W	**Inter-die Bandwidth**	24GB/s
Energy Efficiency (1 core)	20GOPS/W (51pj/OP)	**Expansion Mode**	core-core / core-mem./ core-acc.

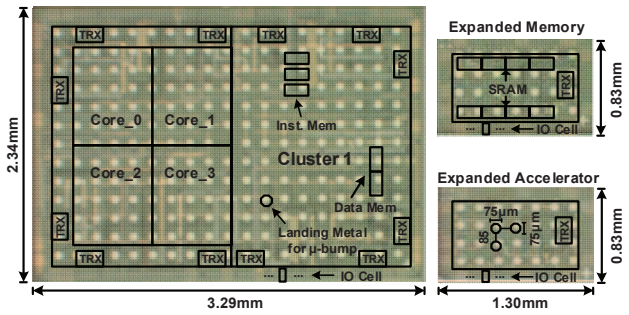

Fig. 7. Chip micrograph and characteristics

This work was supported in part by grant from Samsung Corporation.

References

[1] M. Horowitz. "Computing's energy problem (and what we can do about it)," *IEEE ISSCC Dig. Tech. Papers*, pp. 10-14, Feb 2014.

[2] Y. Deng, and W. P. Maly, "Interconnect characteristics of 2.5-D system integration scheme," *ACM, Proceedings of the 2001 international symposium on Physical design*, pp. 171-175, Apr 2001.

[3] N. Kim, D. Wu, D. Kim, A. Rahman, and P. Wu, "Interposer design optimization for high frequency signal transmission in passive and active interposer using through silicon via (TSV)," *IEEE, Electronic Components and Technology Conference (ECTC)*, pp. 1160-1167, May 2011.

[4] N. Sturcken, E. O'Sullivan, N. Wang, *et al.*, "A 2.5 D integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer delivering 10.8A/mm2," *IEEE ISSCC Dig. Tech. Papers*, pp. 400-402, Feb 2013.

[5] P. T. Huang, L. C. Chou, T. C. Huang, S. L. Wu, *et al.*, "2.5 D heterogeneously integrated bio-sensing microsystem for multi-channel neural-sensing applications," *IEEE ISSCC Dig. Tech. Papers*, pp. 320-321, Feb 2014.

[6] P. Ou, J. Zhang, H. Quan, Y. Li, *et al.*, "A 65nm 39GOPS/W 24-core processor with 11Tb/s/W packet-controlled circuit-switched double-layer network-on-chip and heterogeneous execution array," *IEEE ISSCC Dig. Tech. Papers*, pp. 56-57, Feb 2013.

[7] Z. Yu, K. You, R. Xiao, H. Quan, *et al.*, "An 800MHz 320mW 16-core processor with message-passing and shared-memory inter-core communication mechanisms," *IEEE ISSCC Dig. Tech. Papers*, pp. 64-66, Feb 2012.

[8] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," *ACM, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 169-178, Aug 2000.