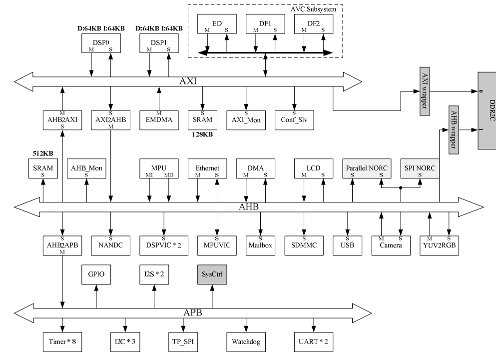
Low Power and High Performance 3-D Multimedia Platform

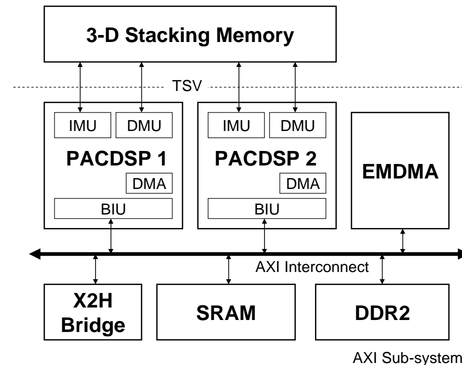
Po-Han Huang, Chi-Hung Lin, Hsien-Ching Hsieh, Huang-Lun Lin and Shing-Wu Tung
Information and Communications Research Lab.
Industrial Technology Research Institute
Hsinchu, Taiwan
E-mail: pohan@itri.org.tw

Traditional technology scaling of semiconductor chips followed Moore's Law. However, the transistor performance improvement will be limited, and designers will not see doubling of operating frequency every two years. Recently, 3-D integrated circuits that utilize through silicon via (TSV) for interconnection have been developed as an improved alternative to the Package-on-Package (PoP) and System-in-Package (SiP) packages. There are many benefits by using TSV-based 3-D integration technologies: (1) Circuit delay can be improved due to the shorter interconnect and reduced parasitic capacitance/inductance, (2) more functionality can be integrated into a small silicon space for form factor reduction and higher packing density due to the additional third dimension, (3) different components with incompatible manufacturing process (i.e. Logic, DRAM, Flash, etc) can be combined in a single 3-D IC for heterogeneous integration. The 3-D integration based on TSV technology enables stacking of multiple memory layers to obtain higher bandwidth for the recent multimedia applications at lower energy consumption. Intel has demonstrated through the teraflops microprocessor chip which is an 80-core design with memory-on-logic architecture. And, each core connects to a 256KB SRAM with 12GB/s bandwidth. Although 3-D IC overcomes many limitations and drawbacks on 2-D IC design, it still has many challenges and design issues that should be considered carefully. In general, the number of TSV is the most critical constraint while designing a 3-D architecture because it is highly related to system performance.

This poster presents a 3-D multimedia platform – 3D-PAC designed by ITRI. 3D-PAC is developed by stacking original 2D-PAC with the SRAM tier. Based on this 3-D stacking technology, the performance can be enhanced about nearly 54% according to different applications. And this poster will show the method of architecture exploration for 3-D stacking. It also describes the detail implementation of reconfigurable SRAM and tier selection when multi-layer stacking SRAM is needed. Finally, the chip is fabricated in TSMC 90nmG CMOS technology. 2D-PAC has the novel features which are described as follows. It is a heterogeneous multi-core architecture, composed of an ARM926EJ-S and two PACDSPs (variable-length, 5-way VLIW architecture designed by ITRI/ICL). This system also consists of three different kinds of buses: AXI, AHB and APB. There are also many peripherals implemented in the system such as I^2C, UART, …, etc.



In general, the new architecture evaluation for an optimized stacking static memory is driven by area, performance, energy efficiency, number of TSVs and thermal issues. After architecture exploration using the electronic system level (ESL) design, the stacking memory is integrated with the instruction memory unit (IMU) and data memory unit (DMU) of PACDSP core because of the performance and number of TSV to build up so called 3D-PAC platform.



For this architecture, each PACDSP owns its private SRAM block (256KB) which is reconfigurable. It allows programmers to configure the different architectures of stacking memories for different applications. For example, a user can configure a part of memory as instruction memory and the rest as data memory or make the entire memory as data memory, which provides high flexibility for the original architecture. With the 3-D stacking SRAM, it equally extends the instruction cache and data memory size of PACDSP. That means programmers can profit in two ways: (1) Reduce the latency caused by cache misses with increasing the size of instruction cache, (2) reduce the frequency of external memory accesses with increasing the size of local data memory.

Two real applications, multi-channel H.264 decoder and JPEG decoder, are chosen to analyze the impact and efficiency from extending PACDSP local memory by 3-D stacking. Experiments are executed on the ESL platform mentioned before. In Multi-channel H.264 decoder application, it takes two PACDSPs to decode four different films simultaneously and display on LCD screen at the same time. It requires more data movement and computation power compared to the

single film decoding.

Experiment configurations (H.264 decoder):

- ARM9 = 204MHz, AHB = 102MHz
- PACDSP = 204MHz, AXI = 204MHz
- DDR2 data rate = 408MHz
- Bitstream:
  SHISEIDO_track1,
  Jomo_track2,
  Ice_Age,
  STC_TEST_Motion
  (10 frames, QVGA)

Experimental results show that overall performance can improve from 11.56fps to 14.80fps (28.02%) without applying any parallelism and optimization for H.264 decoder application. Each DSP is in charge of the decoding of two films. Without 3-D stacking memory, PACDSP needs to backup the internal data to external DDR2 memory during the film switching because of lack of internal data memory, and it incurs a huge overhead. By contrast, with enough 3-D stacking memory supporting (each 256KB), PACDSP does not have to backup related data during film switching. So that will make a huge improvement for system performance. After applying some parallelism and optimization to multi-channel H.264 decoder, the system performance can reach 26.09fps (54.19% improvement compared with 2D 16.92fps).

H.264 Version	Total Cycle	FPS	Improve.
Non-Parallel Version			
2D	194,019,671	11.56	-
3D with 4 binary	151,524,469	14.80	28.02%
Parallel Version			
2D	164,531,751	13.64	-
3D with 2 binary	121,891,167	18.40	34.89%
Parallel Version (Enhanced)			
2D	132,557,335	16.92	-
3D with 2 binary	86,001,705	26.09	54.19%

Experiment configurations (JPEG decoder):

- ARM9 = 204MHz, AHB = 102MHz
- PACDSP = 204MHz, AXI = 204MHz
- DDR2 data rate = 408MHz
- Bitstream: test_image (1 frame, QCIF)

Experiment result shows that there is only little system improvement by extending the instruction cache for JPEG decoder because of the low ratio of cache misses (0.02%). It stands that the original 64KB cache is enough for this application. By contrast, there is huge performance improvement by enlarging the local data memory for JPEG decoder because of the high ratio of external accesses (85.87%). It means the original 64KB data memory is not enough for this application. According to this analysis, it seems that suitable memory configuration depends on different applications. 3D-PAC platform maintains this design feature. Each PACDSP owns its private SRAM block (256KB) which is reconfigurable. It allows programmers to configure the different architectures of stacking memories for different applications.

Execution Type	Cycle Count	Ratio
Cache Size : 64KB		
Cache miss	129,314	0.02%
External access	550,498,586	85.84%
DSP computation	90,676,349	14.13%
Total cycle count		641,304,249
Cache Size : 128KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count		641,029,123

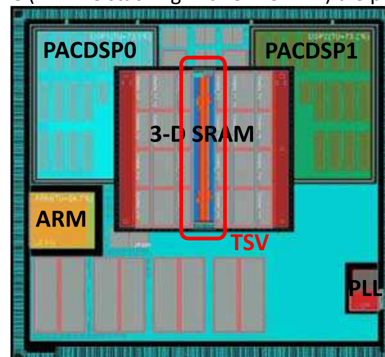
Execution Type	Cycle Count	Ratio
Data Memory Size : 64KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count		641,029,123
Data Memory Size : 64+192KB		
Cache miss	8,865	0.00%
External access	2,051,198	1.96%
DSP computation	102,449,710	98.02%
Total cycle count		104,509,773

There are total 1,914 TSVs in 2D-PAC allocated in the middle area. Related chip SPEC (both 2D-PAC and 3-D stacking SRAM) and die photo shows as follows:

Design	2D-PAC
Process	TSMC 90nmG
Operating Frequency	PACDSP 300MHz
Operating Voltage	Core: 1.0~1.2V I/O: 3.3V
# I/O Pads	498 with PWR/GND
Die Area	7880 x 7880 um^2

Design	3-D stacking SRAM
Process	TSMC 90nmG
Operating Frequency	300MHz
Operating Voltage	Core: 1.0~1.2V
Die Area	3880 x 3880 um^2

3D-PAC (2D-PAC stacking with 3-D SRAM) die photo:



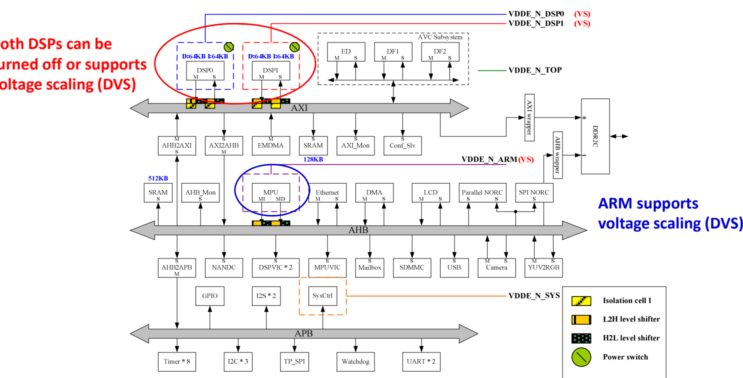
Low-Power and High-Performance 3-D Stacking Multimedia Platform

Industrial Technology Research Institute (ITRI), Taiwan

Po-Han Huang
Chi-Hung Lin
Hsien-Ching Hsieh
Huang-Lun Lin
Shing-Wu Tung

- ■ A low-power & high-performance 3-D stacking multimedia platform is proposed
  - Heterogeneous (logic & memory) integration via 3-D technology
  - Extensive architecture exploration with ESL simulation
- ■ 54.19% performance improvement over 2-D architecture
- ■ Reconfigurable stacking SRAM improves the programming flexibility, where the performance can be optimized for different applications
  - Overhead is neglectable
- ■ Fabricated in the TSMC 90nm generic CMOS technology
  - 1,914 TSVs have been utilized for 3D stacking

2D-PAC with various low-power techniques

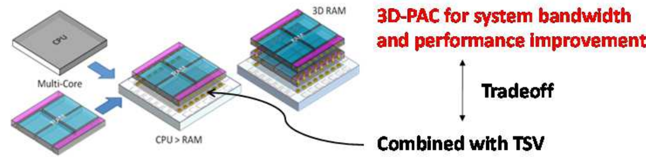


Heterogeneous multi-core architecture:

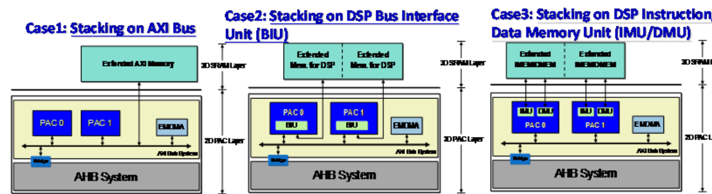
- ☐ ARM926EJ-S
- ☐ Two PACDSPs (variable-length 5-way VLIW)

Low power architecture design flow:
Common Power Format (CPF) by Cadence Inc.

Integration of 3D stacking technology for 3D-PAC



Architecture exploration for 3D-PAC



Performance/Cost Evaluation with ESL

Configurations:

- ☐ ARM9 = 204MHz, AHB = 102MHz
- ☐ PACDSP = 204MHz, AXI = 204MHz
- ☐ DDR2 data rate = 408MHz
- ☐ Application: H.264 decoder
- ☐ Bitstream: foreman (30 frames, QCIF)

Architecture	Total Cycle	FPS	Improve.	# TSV
2D-PAC (DDR2)	164,531,751	13.64	-	-
3D-PAC (3-D SRAM on AXI bus)	162,312,919	13.83	1.39%	272
3D-PAC (3-D SRAM on BIU)	162,308,407	13.83	1.39%	544
✓ 3D-PAC (3-D SRAM on IMU/DMU)	121,891,167	18.40	34.89%	1,886

Configurations:

- ☐ ARM9 = 204MHz, AHB = 102MHz
- ☐ PACDSP = 204MHz, AXI = 204MHz
- ☐ DDR2 data rate = 408MHz
- ☐ Application: JPEG decoder
- ☐ Bitstream: test_image (1 frame, QCIF)

Case1: Cache size↑

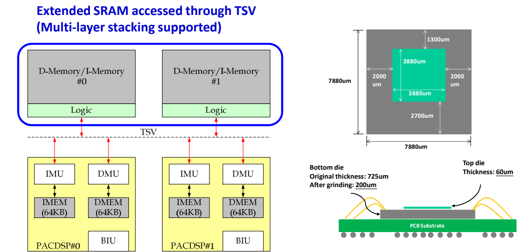
Execution Type	Cycle Count	Ratio
Cache Size: 64KB		
Cache miss	129,314	0.02%
External access	550,498,586	85.84%
DSP computation	90,676,349	14.13%
Total cycle count	641,304,249	
Cache Size: 128KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count	641,029,123	

Case2: Data memory size↑

Execution Type	Cycle Count	Ratio
Data Memory Size: 64KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count	641,029,123	
Data Memory Size: 64+192KB		
Cache miss	8,865	0.00%
External access	2,051,198	1.96%
DSP computation	102,449,710	98.02%
Total cycle count	104,509,773	

Suitable memory configuration depends on different application => Reconfigurable 3-D stacking memory

Final architecture of 3D-PAC



Design SPEC:

- ☐ Private SRAM block (256KB) for each DSP
- ☐ Reconfigurable SRAM as internal instruction or data memory of DSP

Implementation Results

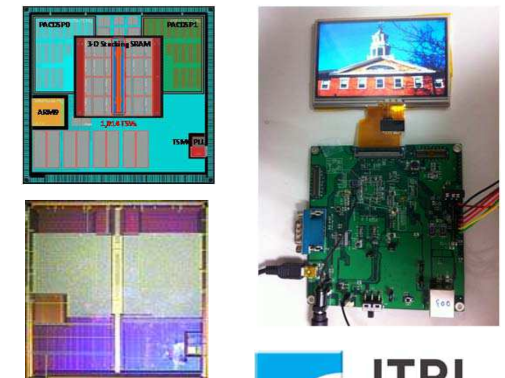
Logic Layer

Design	2D-PAC
Process	TSMC 90nmG
Operating Frequency	PACDSP 300MHz
Operating Voltage	Core: 1.0~1.2V I/O: 3.3V
# I/O Pads	498 with PWR/GND
Die Area	7880 x 7880 um²

Memory Layer

Design	3-D stacking SRAM
Process	TSMC 90nmG
Operating Frequency	300MHz
Operating Voltage	Core: 1.0~1.2V
Die Area	3880 x 3880 um²

Heterogeneous integration with total 1,914 TSVs



ITRI
Industrial Technology Research Institute