# A 3D Stacked High Performance Scalable Architecture for 3D Fourier Transform

George R. Voicu, Marius Enachescu, Sorin D. Cotofana
Delft University of Technology, The Netherlands
{G.R.Voicu, M.Enachescu, S.D.Cotofana}@tudelft.nl

***Abstract*—This paper proposes and evaluates a novel high-performance systolic architecture for 3D Fourier Transform specially tailored for 3D stacking integration with Through Silicon Vias. Our cuboid-shaped systolic network of orthogonally connected processing elements makes use of the DFT algorithm to compute an $N_1 \times N_2 \times N_3$-point 3D-FT with an asymptotic time complexity of $O(N_1+N_2+N_3)$ multiplications. When compared with state-of-the-art 3D-FFT implementation on the Anton machine, a physical synthesized implementation of our architecture on the same 90nm technology node achieves 7.73x and 5.88x speed improvement when computing $16\times16\times16$ and $32\times32\times32$ FT, respectively.**

## I. Introduction

3D Fourier Transform (3D-FT) is employed to reduce the asymptotic computation complexity in numerical simulations of physical or (bio-)chemical phenomena. The cubic growth of the data volume for 3D-FT demands a scalable parallel algorithm with an appropriate implementation. By far the most used approach for solving an $N$-point FT is the Fast Fourier Transform (FFT), with a computational complexity in the order of $O(N \log N)$. A multi-dimensional FT is solved by decomposing it in multiple 1D-FTs, and high performance 3D-FT solutions map efficient FFT algorithms on a cluster of general purpose or specialized computing nodes [1].

A suitable alternative to accelerate the demanded multi-dimensional computations is to use multi-dimensional systolic structures, adequate for CMOS integration [2]. 3D systolic structure for matrix multiplication suited for physical 3D layouts were proposed in [3] while in [4] perimeter I/O on planar systolic arrays for image processing are replaced with 2D area I/O streams. To the best of our knowledge little or no effort has been spent afterwards to map systolic structure on physical 3D layouts, most likely because the required technology to manufacture true 3D ICs, i.e., Through Silicon Vias (TSV), has only recently become competitive [5].

Hardware implementations of 1D High Radix FFTs have been mapped in 3-tiers stacked solution [6], [7]. Although gains in speed and energy efficiency are obtained due to wirelength shortening when the centralized memory is distributed across tiers, the approach simply optimizes the 2D-layout architecture for the stacked-layout case. However, the third vertical dimension introduced by the stacking of silicon tiers changes the way a general-purpose or specific micro-architecture should be designed. Physical arrangements with large number of blocks and high interconnect density between blocks are easier and more efficient to build when the physical design space is extended to three dimensions.

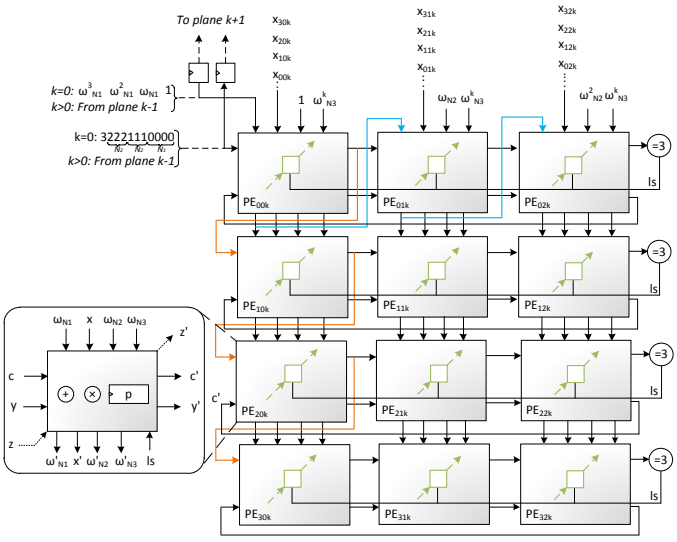


Fig. 1: Layout of a $4 \times 3$ horizontal plane of the 3D systolic network.

In this paper we envision an architecture designed from ground up to exploit the benefits of 3D stacked integration. We propose a high performance cuboid-organized systolic network with $N_1 \times N_2 \times N_3$ orthogonally connected Processing Elements (PEs). A 3D $N_1 \times N_2 \times N_3$ FT is computed by our systolic architecture with an asymptotic time complexity of $O(N_1+N_2+N_3)$ multiplications, by decomposing the 3D-FT in three successive one-dimension Discrete FT (DFT). The physical layout contains $N_3$ identical tiers, and TSVs with a uniform distributed density connect PEs in adjacent tiers.

Compared with the fastest computation of 3D-FT currently available, our synthesized implementation in the same 90nm CMOS technology node achieves 7.73x and 5.88x speed-up when computing $16\times16\times16$ and $32\times32\times32$ FT, respectively.

## II. 3D-Stacked systolic architecture for 3D FT

To solve an $N_1 \times N_2 \times N_3$ 3D-FT we propose a high performance systolic network DFT architecture organized as a rectangular cuboid of orthogonally connected identical PEs. We built an $N_3$-tier stack of horizontal layers containing $N_1 \times N_2$ PEs (Fig. 1) vertically connected through $z$ and $z'$ signals. In each silicon layer we extend the PEs and the control flow in the 2D systolic array for 2D-DFT from [8] to route the results of 2D-FT back in the network and compute 3D-FT.

The processing in the 3D-DFT systolic network is split into four stages, with the corresponding actions and data flow patterns from Table I. The utilized algorithm is the common

TABLE I: PROCESSING ELEMENT ACTIONS AND NETWORK DATA FLOW

| Stage | $S_1$ (c=0) | $S_2$ (c=1) | $S_3$ (c=2 & ls=0) | $S_4$ (ls = 1) |
|---|---|---|---|---|
| Actions | $x' \leftarrow \omega_{N_1} \cdot x$<br>$p \leftarrow p + x$<br>$\omega'_{N_1} \leftarrow \omega_{N_1}$<br>$c' \leftarrow c$ | $y' \leftarrow y + p$<br>$p \leftarrow \omega_{N_2} \cdot p$<br>$\omega'_{N_2} \leftarrow \omega_{N_2}$<br>$c' \leftarrow c$ | $y' \leftarrow p$<br>$p \leftarrow y$<br>$c' \leftarrow c$ | $z' \leftarrow z + p$<br>$p \leftarrow p \cdot \omega_{N_3}$<br>$\omega'_{N_3} \leftarrow \omega_{N_3}$ |
| Data flow | Columns in each tier | Rows in each tier | Rows in each tier | Vertical columns |

dimensional decomposition, in which first $N_2 \times N_3$ 1D-DFTs of size $N_1$ are computed in $S_1$, followed by $N_1 \times N_3$ 1D-DFTs of size $N_2$ in $S_2$, and $N_1 \times N_2$ 1D-DFT of size $N_3$ in $S_4$.

A PE is composed out of a complex multiplier and adder, and registers to shift and hold the complex values of the externally supplied twiddle factors ($\omega'_{N_1}, \omega'_{N_2}, \omega'_{N_3}$) and of the generated intermediate results ($x', y', z'$, and $p$). The activation of each PE and its transition between stages is controlled by the external control vector $c$, and the internally generated signal $ls$. Input data are fed in a diamond shaped form at the top first row of each network layer. At the end of $S_1$ the intermediate result matrix resides in the $p$ registers. The data flow is simply switched in $S_2$, hence no data transposition is needed. In $S_3$ the intermediate results of $S_2$ are looped-back in the network so $S_4$ can switch the data flow to the vertical direction and produce the final results at the $z'$ outputs of the top layer.

Stage $S_1, S_2$, and $S_4$ perform 1D-DFT in $N_1, N_2$, and $N_3$ time steps, respectively, and $S_3$ takes $N_2$ time steps to complete. The time step is defined as the maximum propagation delay of a PE operation, which is dominated by the complex multiplier latency. Since all four stages are pipelined, when computing the total network delay we must also consider the initial computation propagation in all cells across the network diagonal. The systolic network computes the 3D-DFT in $2N_1 + 3N_2 + 2N_3 - 3$ time steps, thus it performs a full $N_1 \times N_2 \times N_3$ 3D-FT with an asymptotic time and area complexity in the order of $O(N_1 + N_2 + N_3)$ complex multiplications, and $O(N_1 \times N_2 \times N_3)$ PEs, respectively.

## III. PERFORMANCE EVALUATION

We compare our proposal with the fastest solution for 3D-FT, namely an FFT implementation on the Anton machine [1], a cluster of nodes logically tied together in a 3D spatial torus network. Each of Anton's nodes contain four Tensilica LX cores and eight coprocessors doing the bulk of FFT operations, and is implemented as a planar ASIC. Thus, no 3D stacking is utilized and the inter-node communication is done through on-board links.

Our 3D-FT specialized architecture is also based on a spatial 3D node (PE) arrangement, but it is naturally mapped on a 3D stacked layout. This combined with the fact that the PEs have a simpler functionality (each PE contains a complex multiplier and adder) allows for the entire network to be contained in a single package, leveraging on-chip high bandwidth links.

Execution times for different transform sizes of our implementation based on worst-case timing results after physical synthesis and of Anton are presented in Table II. We use the RC equivalent model from [9] for the TSV. We note that even

TABLE II: 3D-FT EXECUTION TIMES COMPARISON

| Implementation | | Frequency [MHz] | Time [µs] for transform size 16×16×16 | 32×32×32 | 64×64×64 |
|---|---|---|---|---|---|
| Anton (CMOS 90nm, Fixed) | | 485 | 2.4 | 3.7 | 13.2 |
| 3D-Stacked 3D-FT (CMOS 90nm + TSV) | Fixed | 351 | 0.31 | 0.629 | 1.267 |
| | FP | 174 | 0.625 | 1.266 | 2.551 |

* Fixed = Q8.23 fixed point, FP = IEEE754 single-precision floating point

though our architecture has the number of nodes in the systolic network equal with the transform size, the operating frequency is only determined by the computational logic inside the PE, hence it is constant for all transform sizes.

Our implementation achieves 7.73x, and 5.88x speed-up when computing 16×16×16 and 32×32×32 FT, respectively. The duration of the bare arithmetic computations of the three 32-point FFTs on Anton is under 1.5 µs [1], inter-node data transfers accounting for the rest of the time and limiting the performance. This is even more apparent in the 64×64×64 FT case, when the number of resources in the Anton hardware becomes limited for the problem size, hence a larger speed-up of 10.4x is achieved. The smaller gain for the 32×32×32 FT can be explained by the Anton's optimization for that case, while our network is linearly scalable with the transform size.

It is important to note that the FFT algorithm drastically reduces the number of operations (hence also the number of intermediate data transfers between computing elements) when compared to the DFT algorithm. Therefore, we can draw the conclusion that the increase in bandwidth offered by 3D Stacked integration is capable to sustain an increase in performance even when redundant computations are executed.

Further 88% and 105% frequency improvement for fixed and floating-point PEs, respectively, when using a 45nm technology prove the architecture potential for scaling with the reduction of the technology node size.

## IV. CONCLUSIONS

We described and evaluated a high-performance 3D systolic architecture for the computation of 3D Discrete Fourier Transform specially designed to exploit the reduction of total interconnect length offered by the third dimension in 3D stacked integration with TSVs. Results indicate that it outperforms the state-of-the-art 3D-FFT and offers better scaling with regard to the transform size and technology node size reduction.

## REFERENCES


[1] C. Young *et al.*, "A 32x32x32, spatially distributed 3D FFT in four microseconds on Anton," in *SC '09*, Portland, Oregon, USA.

[2] H. Lim and E. Swartzlander, "A systolic array for 2-D DFT and 2-D DCT," in *IEEE ASSAP*, San Francisco, CA, USA, 1994, pp. 123–131.

[3] S. Lakhani *et al.*, "2D matrix multiplication on a 3D systolic array," *Microelectronics Journal*, vol. 27, no. 1, pp. 11–22, 1996.

[4] S. Chai and S. Wills, "Systolic opportunities for multidimensional data streams," *IEEE Trans. Parallel Distrib. Syst.*, pp. 388–398, 2002.

[5] H. Chaabouni *et al.*, "Investigation on TSV impact on 65nm CMOS devices and circuits," in *Proc. IEEE IEDM*, 2010, pp. 6–8.

[6] W. R. Davis *et al.*, "An 8192-point Fast Fourier Transform 3D-IC case study," in *Proc. 51st MWSCAS*, 2008, pp. 438–441.

[7] T. Thorolfsson *et al.*, "Design automation for a 3DIC FFT processor for synthetic aperture radar," in *46th ACM/IEEE DAC*, 2009, pp. 51–56.

[8] S. Sedukhin, "A new systolic architecture for pipeline prime factor DFT-algorithm," in *4th Great Lakes Symp. on VLSI, GLSV*, 1994, pp. 40–45.

[9] G. Katti *et al.*, "3D stacked ICs using cu TSVs and die to wafer hybrid collective bonding," in *IEEE IEDM 2009*, Baltimore, Maryland, USA.