Chapter 11
3D for Efficient FPGA

Zvi Or-Bach

11.1 Historical Prospective

Logic devices have amounted to about two thirds of the IC industry for many years. In logic devices, there has always been a tradeoff between the costs of developing the logic device in time and money, versus the cost of the end product in terms of performance, power, and cost ("PPC") as illustrated in Fig. 11.1.

In a fundamental work at the Berkeley Wireless Research Center and followed work at many other technology centers [1–3] this tradeoff has been characterized over two decades of designs and benchmarks (Fig. 11.2).

At the early days of the FPGA market, two programming technologies were competing—SRAM based Look Up Table (LUT), and Anti-Fuse. LUT eventually won because it allows easy technology scaling and unlimited reprogramming iterations. Yet, due to the severe PPC penalties of FPGA technology [4], the adoption of the FPGA technology remains limited (Fig. 11.3).

Adapting 3D technology to FPGA design could be cost-effective and might greatly reduce those PPC penalties.

11.2 Early Work on 3D FPGA

Early work on 3D FPGA considered that forming the SRAM of the LUT on top of the FPGA logic would be technologically possible and far less demanding than forming two levels of logic one on top of the other. Tier Logic collaborated with Toshiba [5] to build SRAM using Thin Film Transistors (TFT) for the FPGA LUT on top of the rest of the FPGA circuit. It believed it could have reduced the FPGA device area by

Z. Or-Bach (✉)
MonolithIC 3D Inc., 3555 Woodford Dr., San Jose, CA 95124, USA
e-mail: Zvi@MonolithIC3D.com

B. Murmann and B. Hoefflinger (eds.), *NANO-CHIPS 2030*,
The Frontiers Collection, https://doi.org/10.1007/978-3-030-18338-7_11

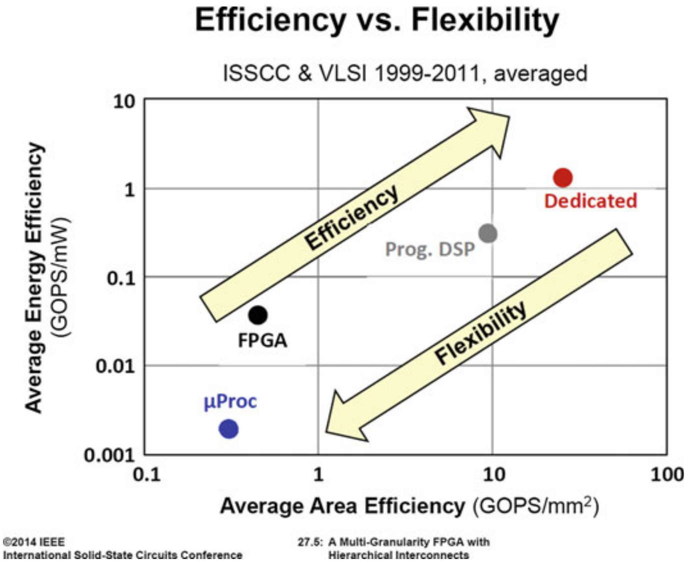


Fig. 11.1 Logic device tradeoff

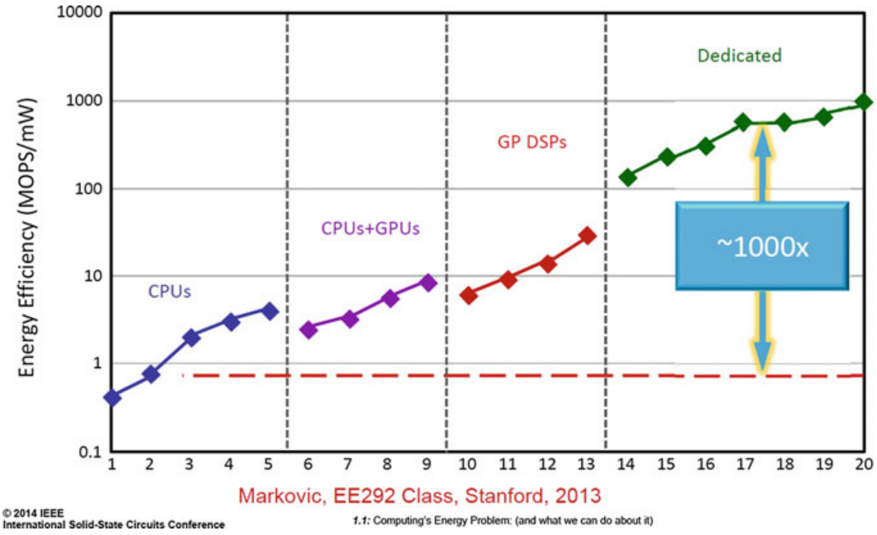
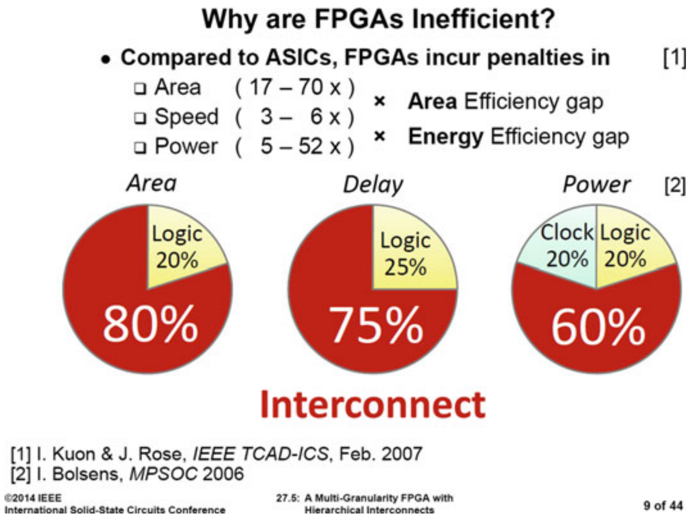


Fig. 11.2 Characterization of logic device tradeoff

about 20%, yet the effort failed, and the project was shut down. A similar concept using RRAM [6] on top of the logic instead of TFT reported potential 40% reduction compared to 2D FPGA but was not pursued commercially.



Why are FPGAs Inefficient?

- Compared to ASICs, FPGAs incur penalties in [1]
  - ❑ Area (17 – 70 x)
  - ❑ Speed (3 – 6 x)
  - ❑ Power (5 – 52 x)

× **Area** Efficiency gap

× **Energy** Efficiency gap

[2]

[1] I. Kuon & J. Rose, *IEEE TCAD-ICS*, Feb. 2007
[2] I. Bolsens, *MPSOC* 2006

©2014 IEEE
International Solid-State Circuits Conference

27.5: A Multi-Granularity FPGA with
Hierarchical Interconnects

9 of 44

Fig. 11.3 The FPGA penalties

CEA Leti has been developing sequential monolithic 3D calling it CoolCube™. As a benchmark, they evaluated [7] applying their technology for FPGA putting logic over memory with the expectation to achieve 55% area reduction compared to 2D FPGA [9].

11.3 3D for Multi-configurations

Tabula, a recently failed start-up, had developed a unique type of FPGA—a real time reconfigurable FPGA. The concept tries to leverage FPGA reconfigurability through storing multiple configurations on-chip and swapping them as needed. It effectively attempted to compensate for the limited area efficiency of the FPGA by reusing the same chip's real estate for multiple purposes on the fly. The company even called its product a 3D FPGA, time being the 3rd dimension. Tabula had raised about $200M but eventually went out of business. An interesting concept that could be added to Tabula structure has been suggested [8] to leverage monolithic 3D technology for multi-stack to hold the multi configuration of the FPGA. Having more than one configuration of a device stack in 3D could allow switching between device configurations within just a few clock cycles and would not increase the device footprint.

11.4 3D for FPGA-ASIC Dual Mode Concept

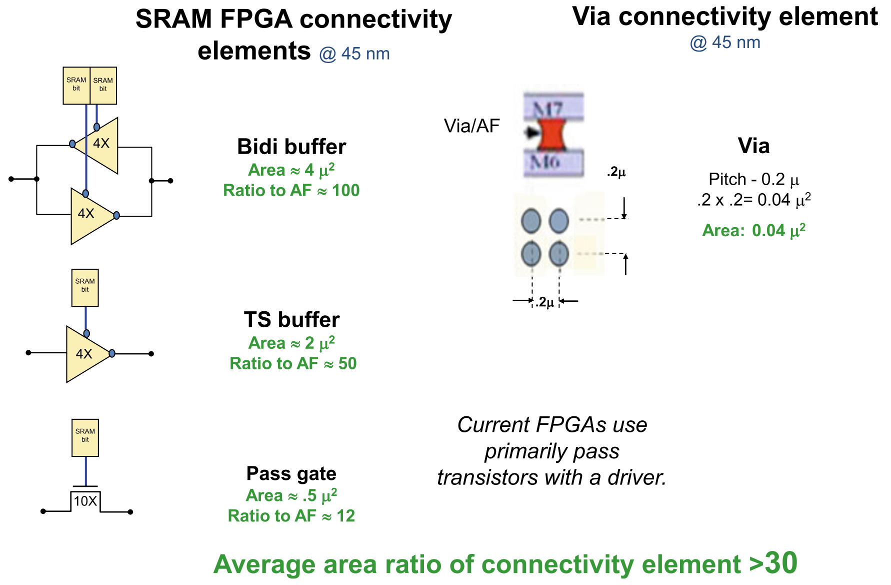
An interesting alternative to FPGA was developed by eASIC [10], recently acquired by Intel. The original concept pioneered by eASIC was that the key deficiency of FPGA is its Programmable Interconnect ("PIC") rather than logic. Consequently, eASIC's early product used programmable LUT-4 (SRAM based) with mask-defined via interconnection. Figure 11.4 illustrates the advantage of via defined interconnect versus PIC at the 45 nm node.

It should be noted that PIC requires sharing some of the base silicon fabric and consumes additional routing resources by going down from the interconnect levels (metal layers 3–6) to the base silicon and up again.

Figure 11.5 illustrates the effectiveness of via-defined interconnect logic. It could potentially provide logic that has only a factor of 2–4 area penalty versus ASICs, with a power-speed penalty of 2–3.

Leveraging monolithic 3D technology could enable effective replacement of eASIC's via with electrically programmable anti-fuse, thus enabling FPGA devices with better than 10× improvement to PPC.

3D heterogeneous integration could help overcome some of the known limitations of anti-fuse technology. First, it allows using a standard fab and process for the base FPGA fabric. Second, it allows saving on the anti-fuse high voltage programming circuits overhead by moving them to an upper level.



SRAM FPGA connectivity elements @ 45 nm

Via connectivity element @ 45 nm

Bidi buffer
Area ≈ 4 μ²
Ratio to AF ≈ 100

TS buffer
Area ≈ 2 μ²
Ratio to AF ≈ 50

Pass gate
Area ≈ .5 μ²
Ratio to AF ≈ 12

Via
Pitch - 0.2 μ
.2 x .2= 0.04 μ²
Area: 0.04 μ²

Current FPGAs use primarily pass transistors with a driver.

Average area ratio of connectivity element >30

Fig. 11.4 Programmable interconnect versus masked defined interconnect

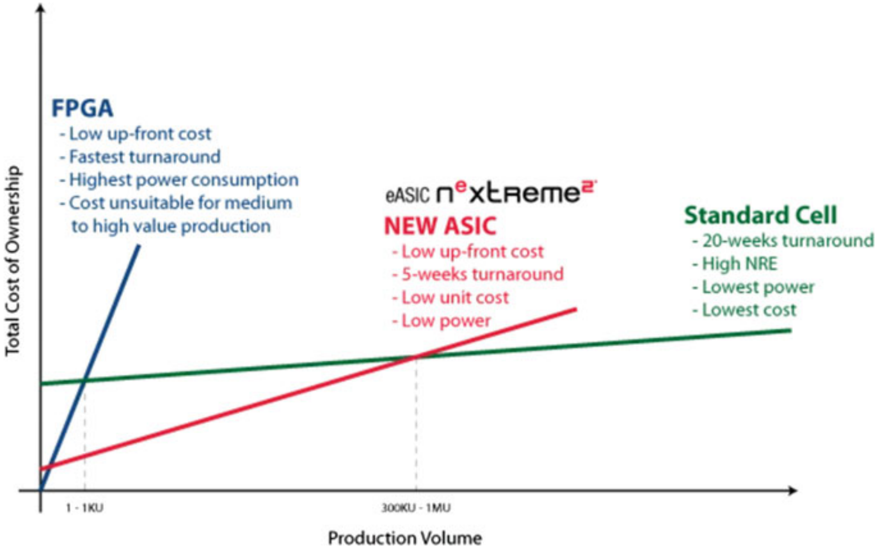


Fig. 11.5 eASIC versus FPGA and versus ASIC. *Source* eASIC web site

Replacing via-defined interconnect fabric with programmable anti-fuse interconnect fabric could be done with relatively low overhead (<20%) as is illustrated by Fig. 11.6.

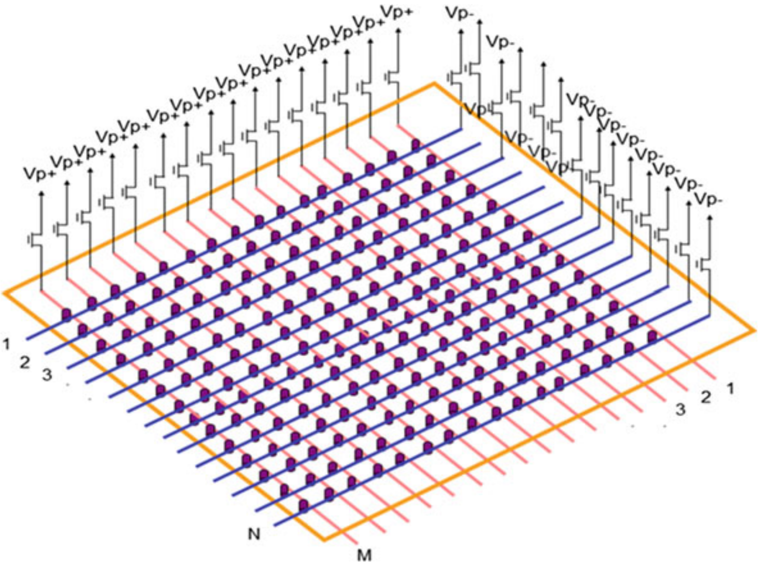


Fig. 11.6 Anti-fuse M × N fully populated crossbar interconnect structure

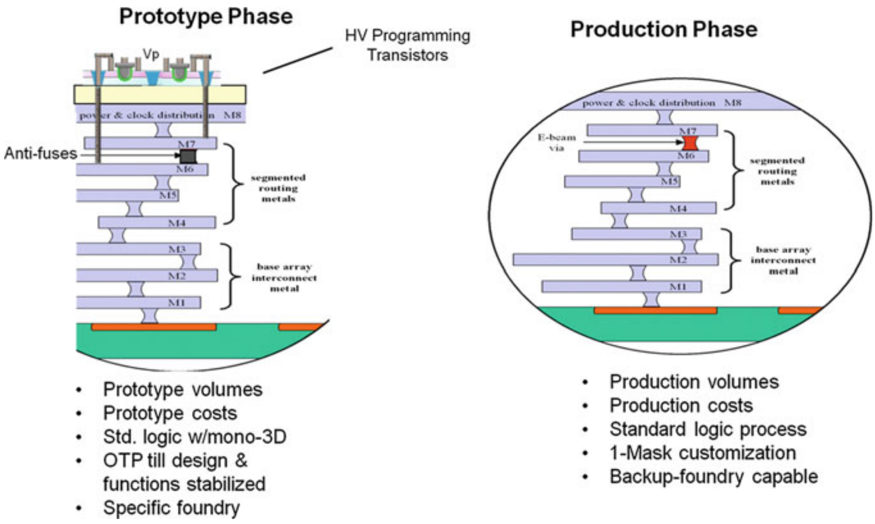


Fig. 11.7 Dual mode: FPGA for prototype and low volume, and mask-defined via for low cost

An additional advantage in which 3D heterogeneous integration could be applied is supporting dual mode of the custom logic: using field programmable device for prototypes and low volume, and form a low-cost compatible volume replacement device, in which the anti-fuses are replaced by a mask-defined via layer (Fig. 11.7).

Removing the anti-fuse and programming circuitry could reduce costs of the high-volume part for the relatively low cost of a single via mask.

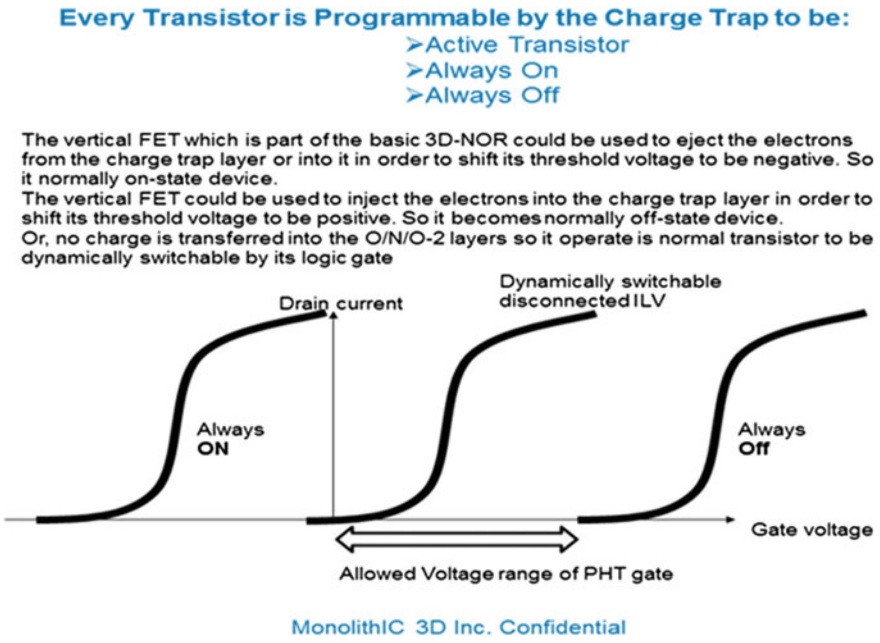
11.5 Utilizing 3D Memory Fabric for FPGA Fabric

The breakthrough which was introduced with 3D NAND technology was the introduction of a new form of scaling—3D Scaling. In 3D scaling technology, more device transistors (or memory cells) are being produced for about the same manufacturing effort by having more layers in the substrate starting wafer. In Chap. 10 we presented a variation called 3D NOR which could be used to replace Stacked Capacitor DRAM technology. Here, a technology concept is presented to leverage 3D scaling for FPGA fabric. The technology has also been detailed in MonolithIC 3D, Inc. patent applications [11, 12]. The first structure [11] is leveraging 3D NOR memory fabric having a single crystal channel and vertically oriented word-lines for FPGA fabric. The second structure [12] leverages 3D NOR memory fabric having poly-crystalline channel and horizontally oriented word-lines for FPGA fabric. The following description is based on the first structure. First, a generic structure is constructed using shared lithography and processing, which later on could be programmed to function as an FPGA.

11.5.1 The Fabric

A key concept leveraging 3D NOR memory structure for FPGA application is using a flash memory for programmable logic applications [13–15] (Fig. 11.8).

A variation of the 3D NOR structure presented in Chap. 10 could include first epitaxial growth of multilayer SiGe over silicon for single crystal channel, or conventional multilayer deposition of polysilicon over oxide as common for 3D NAND. Then, etching the structure, forming rims and valleys takes place (Fig. 11.9).



Every Transistor is Programmable by the Charge Trap to be:

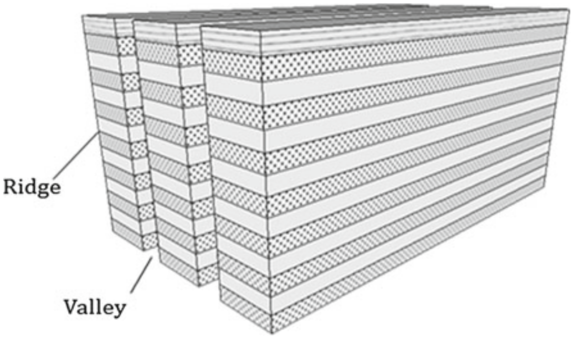
- ➢Active Transistor
- ➢Always On
- ➢Always Off

The vertical FET which is part of the basic 3D-NOR could be used to eject the electrons from the charge trap layer or into it in order to shift its threshold voltage to be negative. So it normally on-state device.
The vertical FET could be used to inject the electrons into the charge trap layer in order to shift its threshold voltage to be positive. So it becomes normally off-state device.
Or, no charge is transferred into the O/N/O-2 layers so it operate is normal transistor to be dynamically switchable by its logic gate

MonolithIC 3D Inc. Confidential

Fig. 11.8 Flash cell is a programmable logic function

Fig. 11.9 Multilayer substrate after etching forming ridges and valleys



Next, depositing Oxide-Nitride-Oxide (O/N/O) makes the structure ready for charge trap memory function. Next, forming gates and a staircase makes the structure illustrated in Fig. 11.10.

The transistor schematic of one ridge is illustrated in Fig. 11.11.

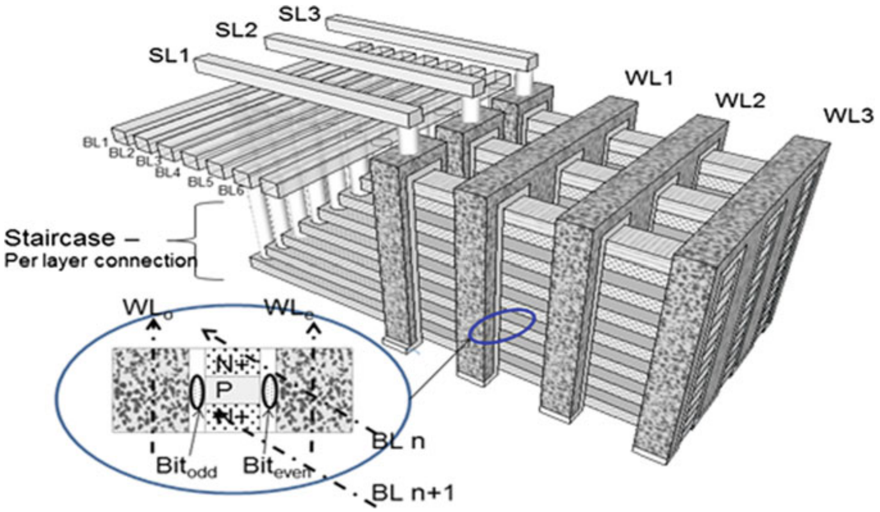
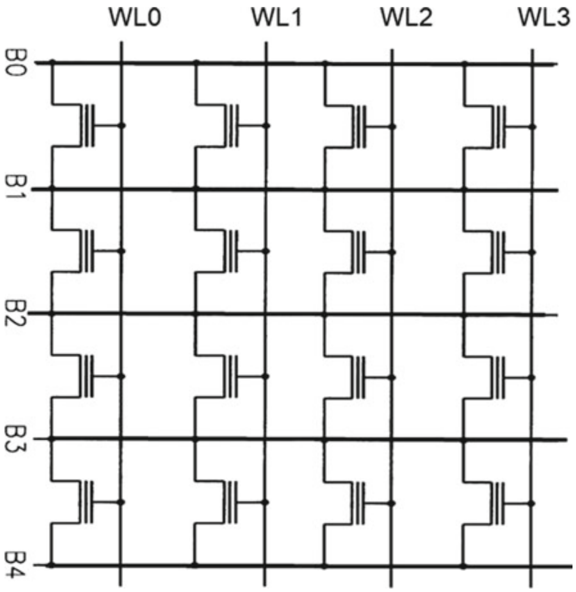


Fig. 11.10 Adding O/N/O, gates, and staircase access

Fig. 11.11 Transistor schematic along a ridge



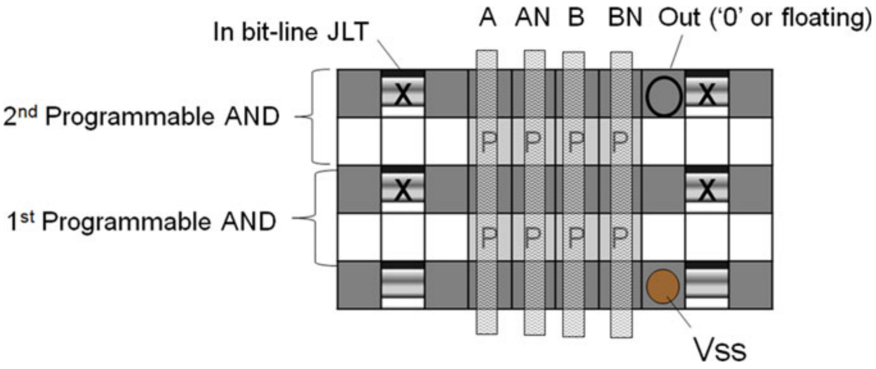


Fig. 11.12 LUT-2 could be formed in section of a 3D NOR structure

11.5.2 Programmable LUT-n Memory

The above structure could be used to form logic functions such as Look-Up-Table and programmable interconnect for FPGA applications. Figure 11.12 illustrates a LUT-2 formed in two layers of such a ridge.

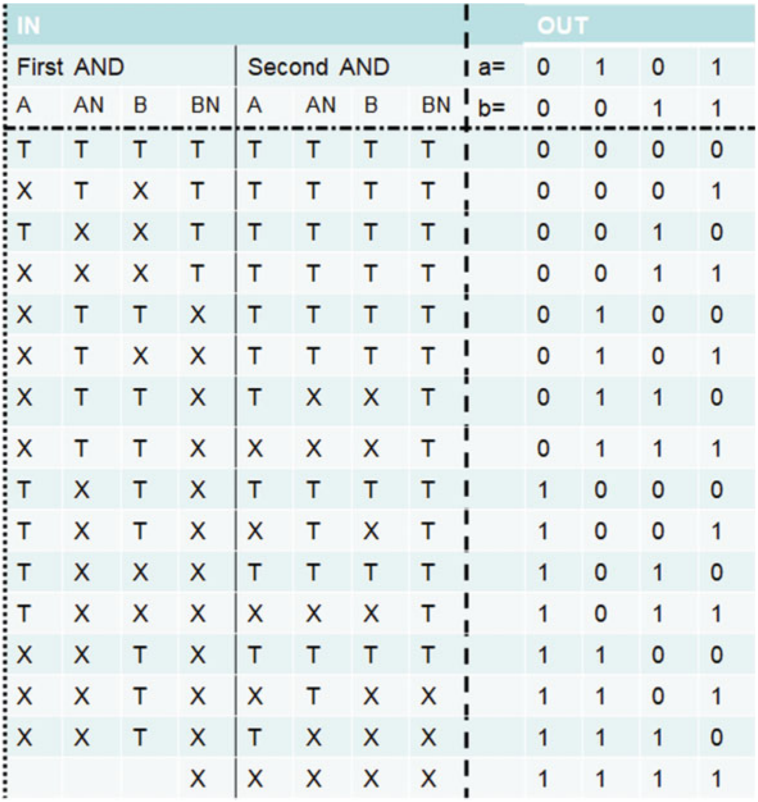
The LUT-2 gates (A, AN, B, BN) are the WL0–WL3 (Fig. 11.11). The X represents an additional variation in which an in the bit-line junction-less-transistors ("JLT") is being formed. The details for such in bit-line JLT processing are detailed in PCT application WO 2017/053329. Such in bit-line JLT enable horizontal segmentation of the 3D NOR structure. The truth table of this LUT-2 structure is presented in Fig. 11.13 (Fig. 11.14).

The 3D NOR structure is a 3D matrix of n-type transistors. Accordingly, the logic functions formed in it utilize only n-type transistors. A transferred layer on top could be used to add full CMOS circuitry to complement the n-only programmable logic underneath. Logic circuits that utilize mainly n-type transistors had been proposed in the past [16]. One approach to reconstruct full swing signals from n-type only circuits is to use two complementing logic functions. Figure 11.15a, b illustrates the use of complementing LUT and LUT-N with top CMOS circuit to reconstruct full swing logic output.

For higher performance, a differential amplifier circuit could be used instead of the logic half-latch.

11.5.3 Programmable Interconnect in Memory

Differential logic could be extended to differential signaling throughout the FPGA. It could help reduce power and improve speed but, far more importantly, it allows using the 3D NOR fabric for programmable routing. Differential interconnects offer lower voltage swings with better noise immunity resulting in lower power. For years,



IN									OUT			
First AND				Second AND				a=	0	1	0	1
A	AN	B	BN	A	AN	B	BN	b=	0	0	1	1
T	T	T	T	T	T	T	T		0	0	0	0
X	T	X	T	T	T	T	T		0	0	0	1
T	X	X	T	T	T	T	T		0	0	1	0
X	X	X	T	T	T	T	T		0	0	1	1
X	T	T	X	T	T	T	T		0	1	0	0
X	T	X	X	T	T	T	T		0	1	0	1
X	T	T	X	T	X	X	T		0	1	1	0
X	T	T	X	X	X	X	T		0	1	1	1
T	X	T	X	T	T	T	T		1	0	0	0
T	X	T	X	X	T	X	T		1	0	0	1
T	X	X	X	T	T	T	T		1	0	1	0
T	X	X	X	X	X	X	T		1	0	1	1
X	X	T	X	T	T	T	T		1	1	0	0
X	X	T	X	X	T	X	X		1	1	0	1
X	X	T	X	T	X	X	X		1	1	1	0
			X	X	X	X	X		1	1	1	1

Fig. 11.13 Truth table of the programmable memory for LUT-2 function

interconnect delay has increased with scaling, while gate delay has decreased as has been illustrated in Fig. 15.2a, b. Yet, the interconnect effect on chip power had been managed by chip operating voltage scaling known as Dennard scaling (Fig. 11.16).

The end of Dennard Scaling made power the limiting factor. The constant charge and discharge of the interconnect capacitance now dominates chip power and performance (Fig. 11.17).

Yet, the industry has not adapted differential interconnect because it requires double the routing resources and additional support circuits. However, as power becomes a dominant problem, perhaps it is time for differential interconnects to take center role in new chip architectures.

3D scaling for configurable logic using shared litho and shared processing opens an iterating opportunity for new type of interconnect technology. In 3D scaling, many layers are processing together, allowing the effective processing of many layers of interconnect together as a generic 3D matrix, and later program them for specific interconnect functions.

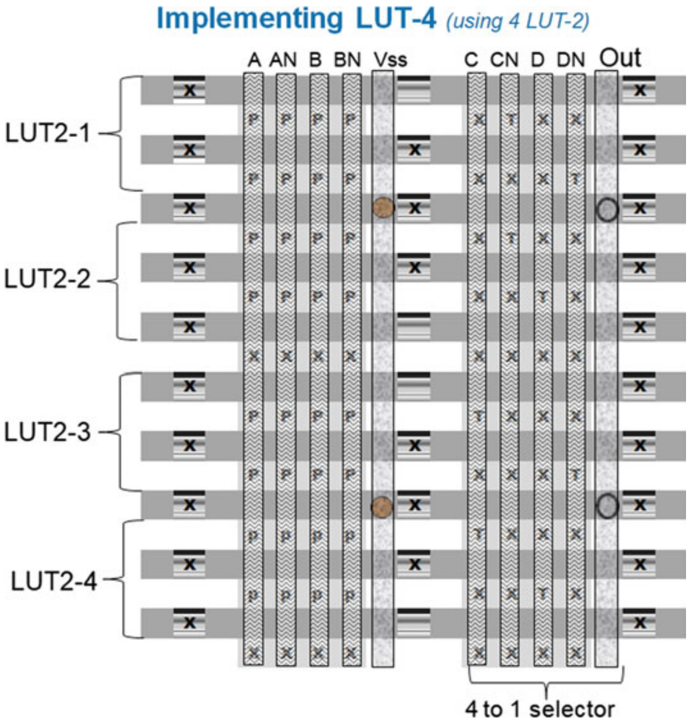


Fig. 11.14 LUT-4 could be formed in section of a 3D NOR ridge structure, having four LUT-2 vertically stacked within a ridge and adjacent 4 to 1 selector

For example, in a 3D fabric of 32 levels the top 10 could be used for the LUT-4 as is illustrated in Fig. 11.14 and the bottom 22 could be used for interconnect. The unused bit-lines of these 22 layers could function as horizontal ("X" direction) segments of the interconnect fabric. Vertical segment could be formed by depositing vertical ("Z" direction) conductive segments in-between the word-lines the structure—see Figs. 11.11 and 11.18a, b.

The programmable connectivity structure could use RRAM technology or anti-fuse (One Time Programmable—"OTP") technology. The connectivity segments in the horizontal direction vertical to the bit-line ("Y" direction), could add in using technology concept know as word-line replacement in 3D NAND (Fig. 11.19).

The support circuit on top could support the differential interconnect just like the differential logic.

The FPGA in memory fabric enables the formation of a multilayer (96–128) memory, such as 3D NOR, with the top 32 layers used for programmable logic while the rest for memory. Recently, logic in memory has become a popular concept as it fits very well many AI type applications. The 3D NOR with built-in FPGA could fit very well in this emerging space.

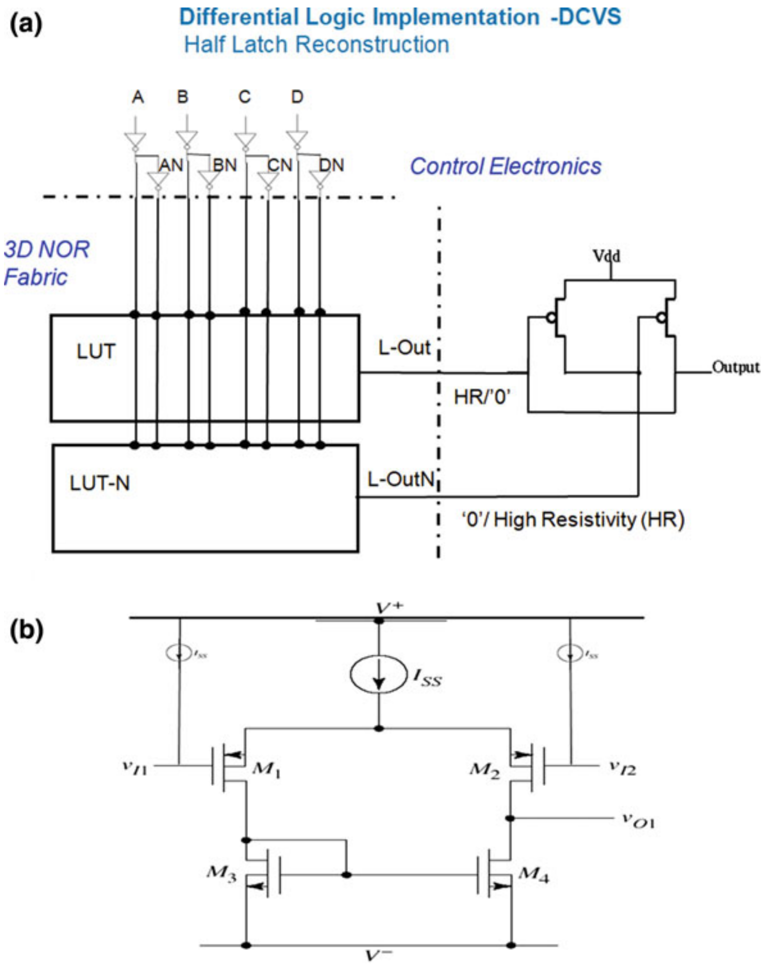


Fig. 11.15 **a** Two complementing LUT-4 with top lower control and reconstruction. **b** Optional differential amplifier top level reconstruction circuit

As a standalone FPGA product, 3D-NOR base FPGA could compete well with mask-defined standard cell designs. The LUT-4 footprint could be about (10 × 100 nm) × (2 × 100 nm) = 0.2 μm^2 which represents a logic density of about 70 MGate/mm^2. The forecast for standard cells at the 7 nm node is about 20 MGate/mm.

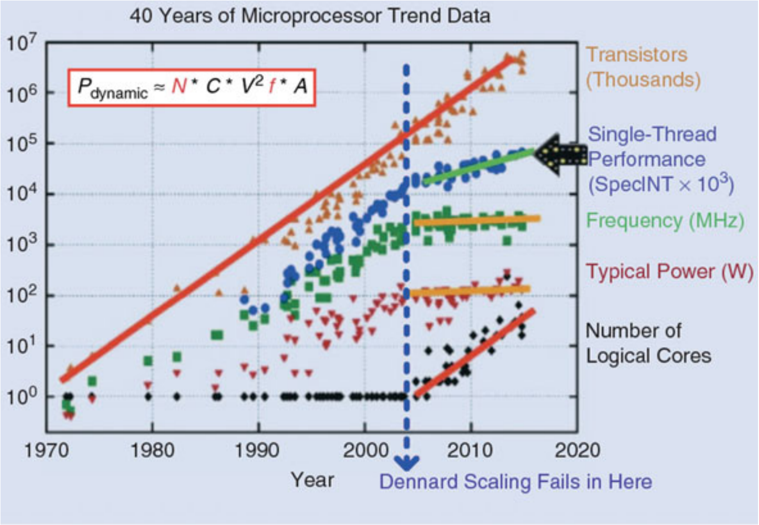
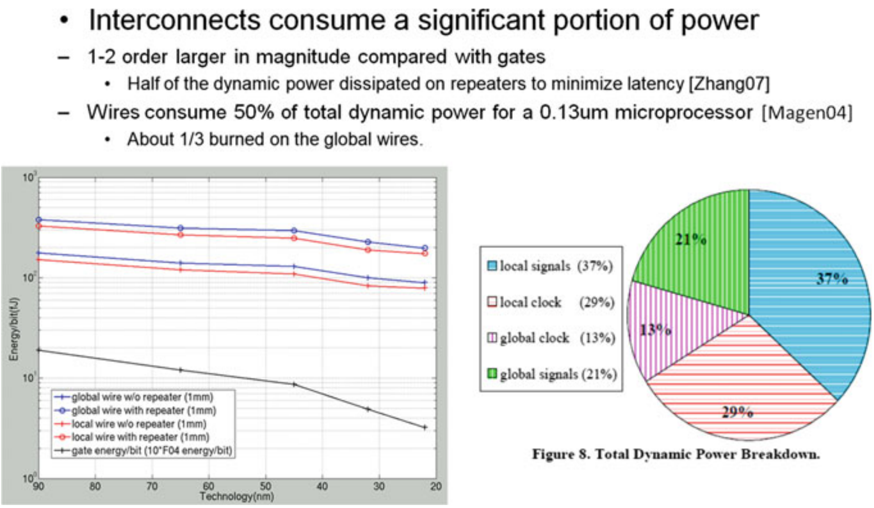


FIGURE 1: The Dennard scaling failed around the middle of the 2000s [24].

Fig. 11.16 End of Dennard scaling [17]



- Interconnects consume a significant portion of power
  - 1-2 order larger in magnitude compared with gates
    - Half of the dynamic power dissipated on repeaters to minimize latency [Zhang07]
  - Wires consume 50% of total dynamic power for a 0.13um microprocessor [Magen04]
    - About 1/3 burned on the global wires.

Figure 8. Total Dynamic Power Breakdown.

Fig. 11.17 Interconnect chip power [18]

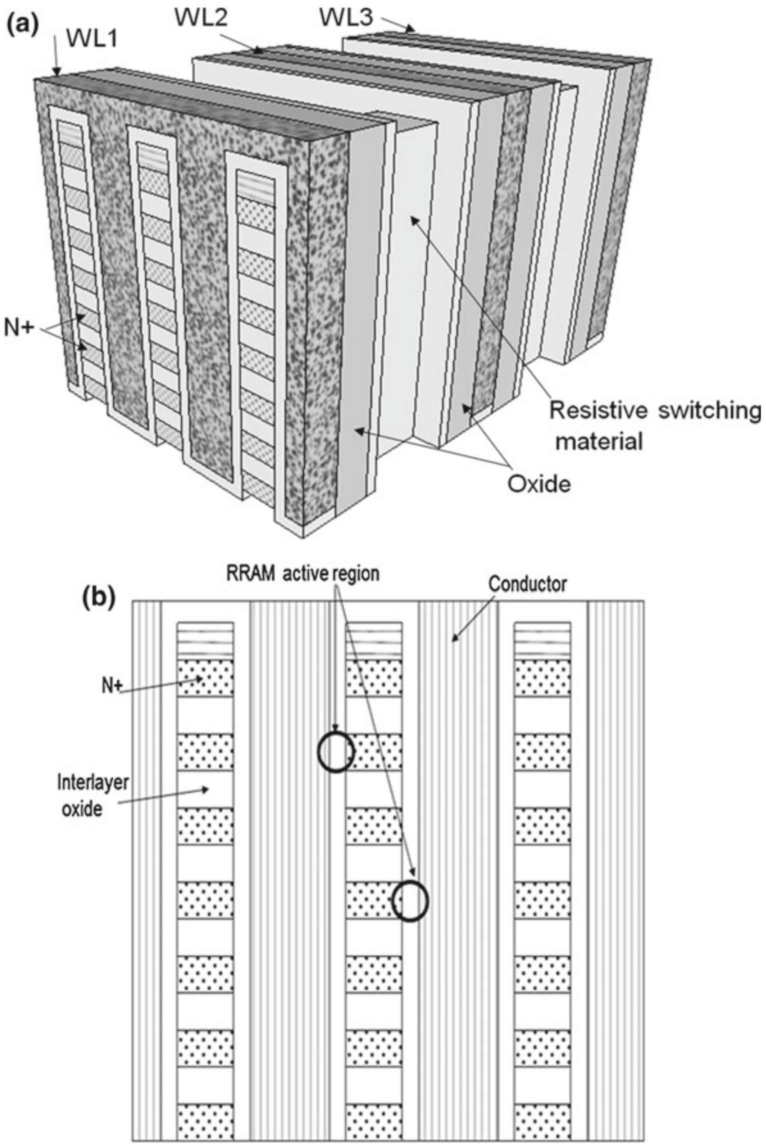


Fig. 11.18 **a** Preparing the structure for Z segments, **b** Z segments with anti-fusses

11.6 Summary

A few alternative concepts have been presented for use of 3D integration in FPGA applications. These alternatives offer different uses of 3D technologies resulting in

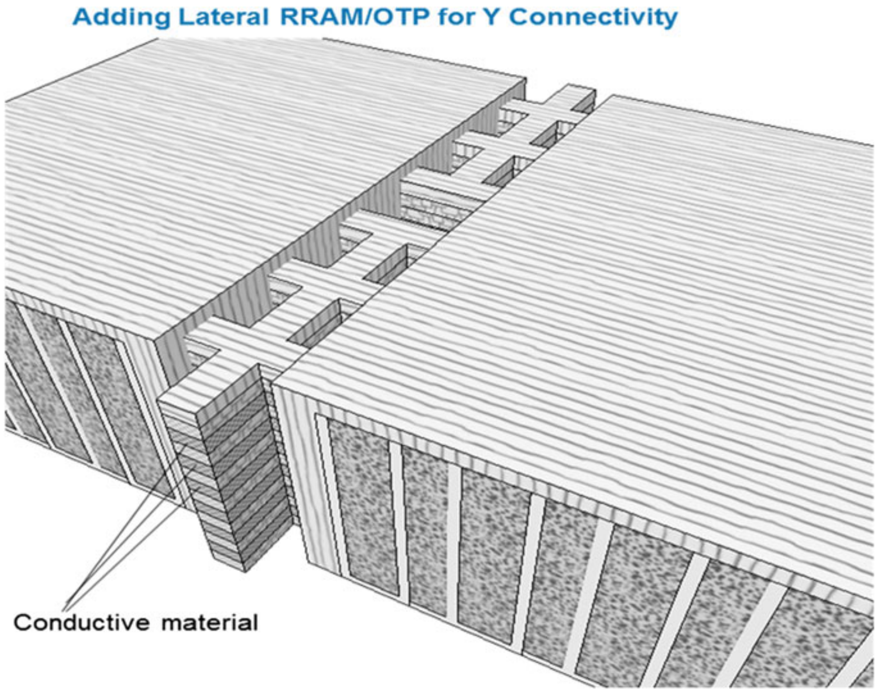


Fig. 11.19 3D structure with programmable logic and X-Y-Z programmable connectivity

different PPC, spanning the spectrum from 2× better FPGA, to about 0.4× of ASIC PPC, and to the 3D NOR FPGA, while having better PPC than ASICs.

References

1. N. Zhang, B. Brodersen, The cost of flexibility in systems on a chip design for signal processing applications. University of California, Berkeley, Tech. Rep. (2002)
2. B. Brodersen, Plenary Session, IEEE S3S 2013
3. M. Horowitz, 1.1 computing's energy problem (and what we can do about it), in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, 2014)
4. F.-L. Yuan et al., A multi-granularity FPGA with hierarchical interconnects for efficient and flexible mobile computing. IEEE J. Solid-State Circuits **50**(1), 137–149 (2014)
5. T. Naito et al., World's first monolithic 3D-FPGA with TFT SRAM over 90 nm 9-layer Cu CMOS, in *2010 Symposium on VLSI Technology* (IEEE, 2010)
6. Y.Y. Liauw et al., Nonvolatile 3D-FPGA with monolithically stacked RRAM-based configuration memory, in *2012 IEEE International Solid-State Circuits Conference* (IEEE, 2012)
7. A. Mihal, S. Teig, A constraint satisfaction approach for programmable logic detailed placement, in *International Conference on Theory and Applications of Satisfiability Testing* (Springer, Berlin, Heidelberg, 2013)
8. US Patent 8,912,820

9. O. Turkyilmaz et al., 3D FPGA using high-density interconnect Monolithic Integration, in *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2014)
10. US Patents: 6,642,744, 6,476,493, 6,819,136
11. Patent Application WO/2017/053329
12. Patent Application WO/2018/144957
13. C. Hu, Interconnect devices for field programmable gate array, in *1992 International Technical Digest on Electron Devices Meeting* (IEEE, 1992)
14. US Patent 5,633,518
15. T. Speers et al., 0.25 μm FLASH memory based FPGA for space applications. System 10000, 100000 (1999): 1000000
16. D. Somasekhar, K. Roy, Differential current switch logic: a low power DCVS logic family. IEEE J. Solid-State Circuits **31**(7), 981–991 (1996)
17. Liming Xiu, Time Moore: exploiting Moore's law from the perspective of time. IEEE Solid-State Circuits Mag. **11**, 39–55 (2019)
18. http://cseweb.ucsd.edu/~kuan/talk/interconnect0824.ppt