Chapter 15
A 1000× Improvement of the Processor-Memory Gap

Zvi Or-Bach

15.1 Historical Prospective

Over more than 50 years, the Integrated Circuit (IC) industry has grown from nothing to over $500 B/year. The driving force was the ability to scale down, known as Moore's Law, where with each new node the number of integrated elements doubles at about the same overall cost and with better speed and lower power. In the deep sub-micron regime such scaling has come at an exponentially higher development and infrastructure cost, usually consisting of many $B. From over 50 IC companies pursuing scaling just 20 years ago, we now have merely three committed to the 7 nm node. Additionally, these handful of companies are integrating just few flavors of logic circuits. Memory circuits are being produced separately by special fabs dedicated to memory. These are DRAM fabs, which at advanced nodes are currently produced by only three vendors, and storage fabs such as 3D NAND. The full system is typically achieved by integrating logic and memory using a Printed Circuits Board (PCB) or 2.5D (chip-on-substrate) packaging. The overall system performance is limited by the off-chip interconnection that lags way behind IC interconnection (Figs. 15.1 and 15.2).

While on-chip interconnects have improved faster than off chip interconnects, they are still far worse than the transistor performance improvement with scaling. And the performance gap between logic gate delay and the on-chip interconnect delay is getting exponentially worse with scaling.

The combination of these effects has been the source of what was called by John L. Hennessy and David A. Patterson the "Memory Wall" [1] or the Processor-Memory Gap. This performance gap has grown by about 50% per year. Figure 15.3a and b shed some more light on this gap.

Z. Or-Bach (✉)
MonolithlC 3D Inc., 3555 Woodford Dr, San Jose, CA 95124, USA
e-mail: Zvi@MonolithlC3D.com; or_bach@yahoo.com

B. Murmann and B. Hoefflinger (eds.), *NANO-CHIPS 2030*,
The Frontiers Collection, https://doi.org/10.1007/978-3-030-18338-7_15

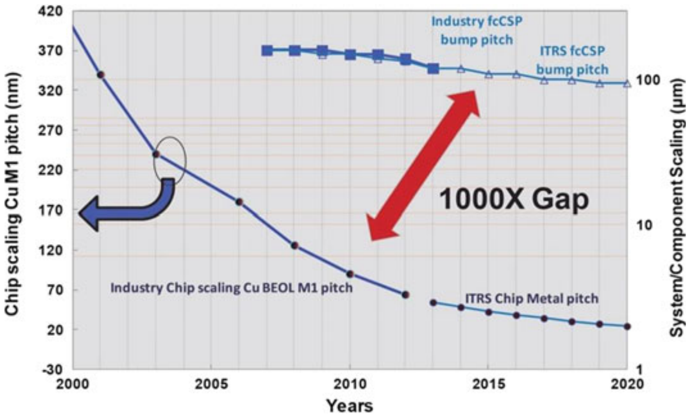


Fig. 15.1 Gap between on-chip interconnect and off-chip interconnect. *Source* VLSI 2013, Dr. Jack Sun, CTO of TSMC

In a report named "Why we need Exascale and why we won't get there by 2020" [3] the problem with the wires has been nicely articulated (see Fig. 15.4).

3D integration leveraging the concepts presented in Chaps. 8 and 10 could help overcoming the memory wall and the tyranny of interconnects to enhance computer systems by orders of magnitude.

The use of Monolithic 3D integration for 1000× improvement in computer performance has been reported [4–6], work on it is now supported in DARPA's 3DSoC program and is also detailed in Chap. 9 of this book (Fig. 15.5).

15.2 Precise Wafer Bonding to Overcome the Memory Wall

The advantage of 3D integration using precise wafer bonders, as detailed in Chap. 8, is the ability to keep using existing wafer processing fabs and processes while allowing 3D heterogeneous integration. Such 3D heterogeneous integration enables overcoming the "Memory Wall" just as suggested in the work by Stanford [4–6].

In a following work [7] the concept of 3D integration has been further advanced to enable first aggregating memory layers, such as conventional DRAM, to create a 3D array of memory with enough capacity and then integrating it with logic to complete the 1000× improved computing system. This concept has been designed to keep the 3D integration as simple as Place-Bond-Thin ("cut")-and Place again. Such simplified 3D integration can leverage Hybrid Bonding [8–11] in which the boding process allows for oxide to oxide and metal to metal bonding, thus achieving both mechanical bonding of the two wafers and formation of electrical connections between the landing pads of the bottom wafer and the connection pins of the top wafers. This could be further enhanced using a technology called "Fusion Hybrid

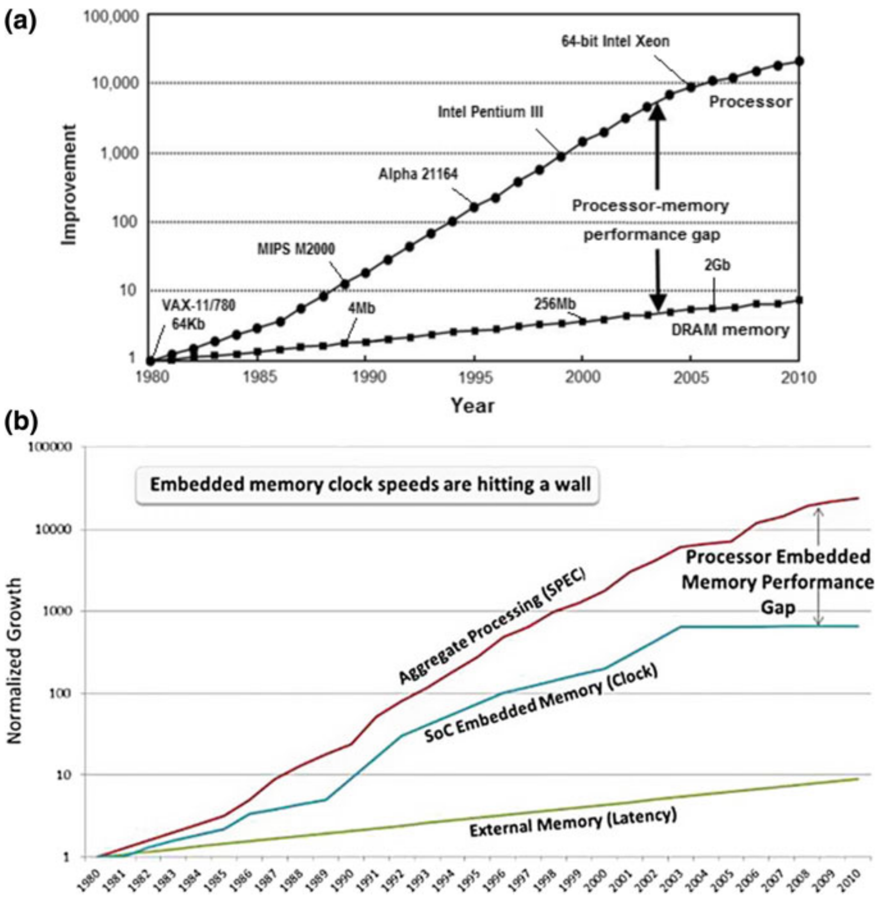
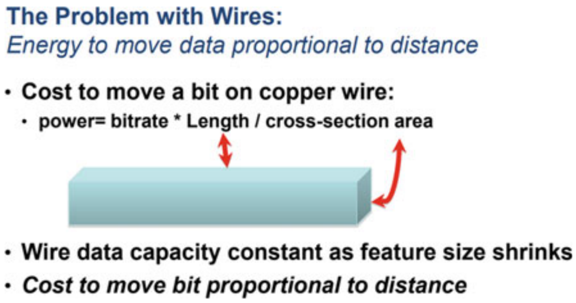


Fig. 15.3 **a** Yearly improvement of processor and DRAM memory speeds over three decades (*Source* [2]). **b** Embedded memory performance gap (*Source* semiwiki.com)

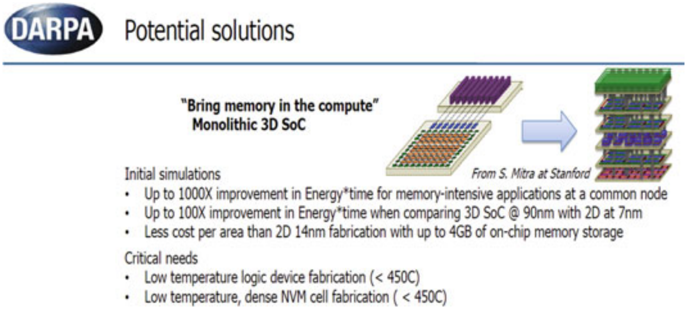
Fig. 15.4 The problem with wires



The Problem with Wires:

Energy to move data proportional to distance

- **Cost to move a bit on copper wire:**
  - **power= bitrate * Length / cross-section area**
- **Wire data capacity constant as feature size shrinks**
- ***Cost to move bit proportional to distance***


DARPA Potential solutions

"Bring memory in the compute"
Monolithic 3D SoC

From S. Mitra at Stanford

Initial simulations
- Up to 1000X improvement in Energy*time for memory-intensive applications at a common node
- Up to 100X improvement in Energy*time when comparing 3D SoC @ 90nm with 2D at 7nm
- Less cost per area than 2D 14nm fabrication with up to 4GB of on-chip memory storage

Critical needs
- Low temperature logic device fabrication (< 450C)
- Low temperature, dense NVM cell fabrication (< 450C)

Fig. 15.5 1000× improvement in energy × time by Monolithic 3D SoC

15.3 The Memory Stack

As presented in Chap. 8, it is desirable to have a ‘cut layer’ built in the base wafer used for the memory stack. Such could be a SiGe layer or an oxide layer or other etch-selective layer. For DRAM wafers the use of the N+ deep well which is common for DRAM wafers could be a convenient option. The use of SOI wafers is also attractive as it allows the use of advanced fab lines such as the GlobalFoundries or Samsung. An additional advantage in the use of SOI, such as GlobalFoundries’ 22FDX process, is having a substrate contact as part of the PDK to provide for back-bias. Such substrate contacts could be used as part of the ‘nano-TSV,’ also called through-layer-via, as illustrated in Figs. 15.6 and 15.7. Vertical pillars are formed with stacking of nano-TSVs.

Use of a ‘cuttable’ wafer enables a controlled removal of the substrate, after its flipping and bonding, by grinding and etching, using the BOX (the ‘cut-layer’) as an etch stop. Accordingly, the ‘nano-TSV’ is made similar to inter-metal via of the corresponding process, which allows about 10,000× higher vertical connectivity [~$(5\mu/50n)^2$]. It should be noted that ‘nano-TSV’ process needs to be all the way to the cut layer, so it could be easily turn into pin or landing pad after flipping, bonding, and cut process, as is illustrated in Fig. 15.7a, b.

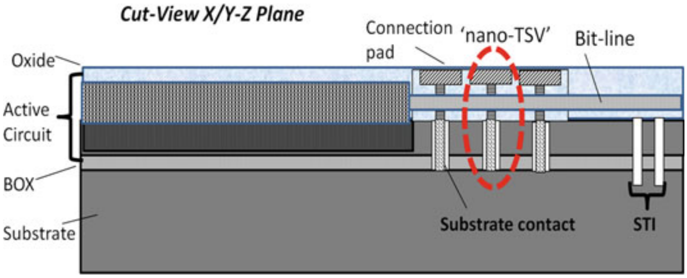


Fig. 15.6 Bit-cell array on SOI wafers with vertical pillar

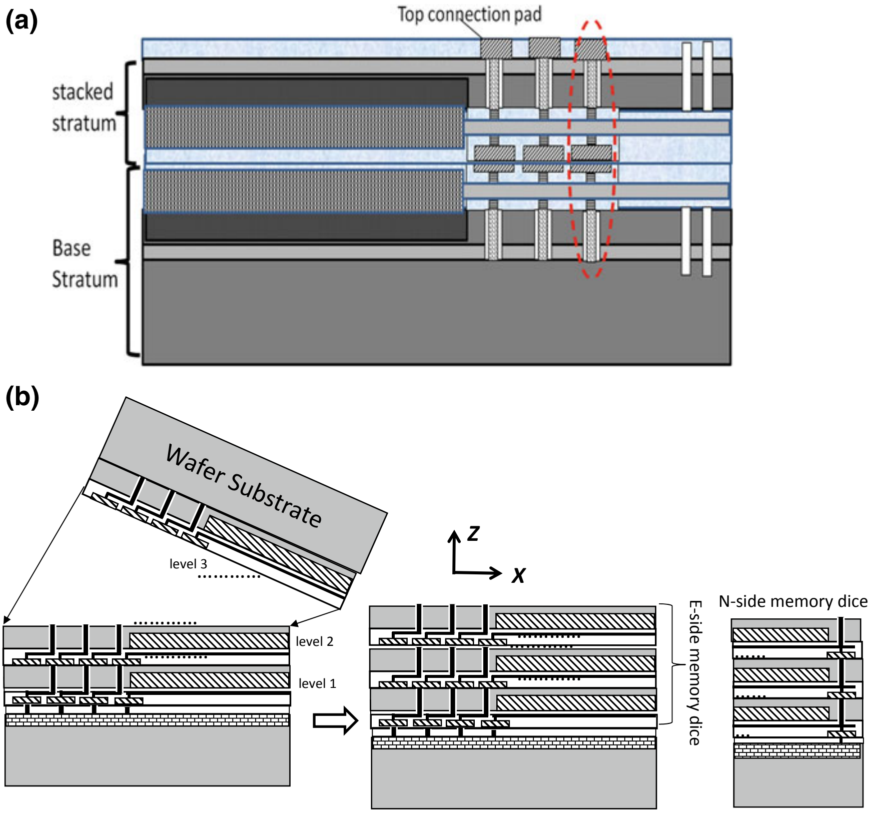


Fig. 15.7 **a** Two memory strata, vertical pillar marked. **b** Illustrating formation flow of three memory strata

The other element enabling fine grain vertical connectivity relates to the stacking misalignment. Until recently, bonder misalignment was on the order of 1 µm, which severely impacted the effective vertical connectivity. To combat that MonolithIC 3D has developed an innovative alignment technique called 'Smart Alignment' [13, 14].

As detailed in Chap. 8 herein, precise bonders are now capable of better than 50 nm (3σ) alignment precision, which removes some of the need for Smart Alignment.

15.4 The Architecture

The suggested computer architecture includes the following strata: Bit-cell array, Memory control, processor, and I/O. Figure 15.5a, b illustrate two optional configurations.

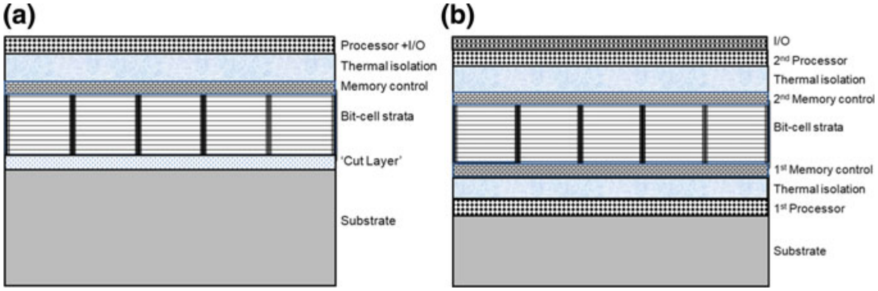


Fig. 15.8 **a** Single side configuration. **b** Dual side configuration

The configuration of Fig. 15.8a is built on a 'cuttable' substrate allowing the use of the illustrated structure as a transferable structure for further 3D integration. The bit-cell memory stack is built by stacking memory strata as will be detailed later. A memory control stratum provides the peripheral circuits for each of the memory units using per unit vertical pillars of global bit-lines and global word-lines. These pillars are formed by stacks of nano-TSVs as illustrated in Figs. 15.6, 15.7. The memory control is interfaced to the processor stratum through a thermal isolation layer designed to isolate the heat generated at the processor stratum from the memory stack underneath it. The processor stratum could include the 3D SoC I/O circuits, or the I/O could occupy its own stratum. Figure 15.8b illustrates an alternative 3D SoC. The base wafer could be any 2D wafer including the most advanced process node for the first processor stratum. Through thermal isolation layer it is connected with the first memory control stratum, which provides bottom peripheral circuits to the memory strata. The memory strata include feed-throughs to allow the bottom side and the top side (2nd memory control) to synchronize their memory access. Overlaying the memory strata is the 2nd memory control stratum, connecting with the 2nd processor stratum built on a 'cuttable' wafer, such as a standard foundry SOI wafer. An I/O stratum overlays the structure, thus providing system connections to the external devices. Such an I/O stratum could be built on a design-rule relaxed SOI process, such as RF-SOI, and could include a wireless communication channel or be built on a wafer supporting optical communication channels.

15.5 Details of the Memory Stack

The memory stack is built by stacking wafers structured as units of bit-cell array [13].

Figure 15.9 illustrates a small 3 × 3 region of an array of units forming the bit-cell array stratum. The unit size is about 200 μm × 200 μm while the connectivity lane between units, intended for inter-stratum connections, is about 1 μm wide (the drawing is not to scale). Each unit is a mini array of tightly packed bit-cells. The

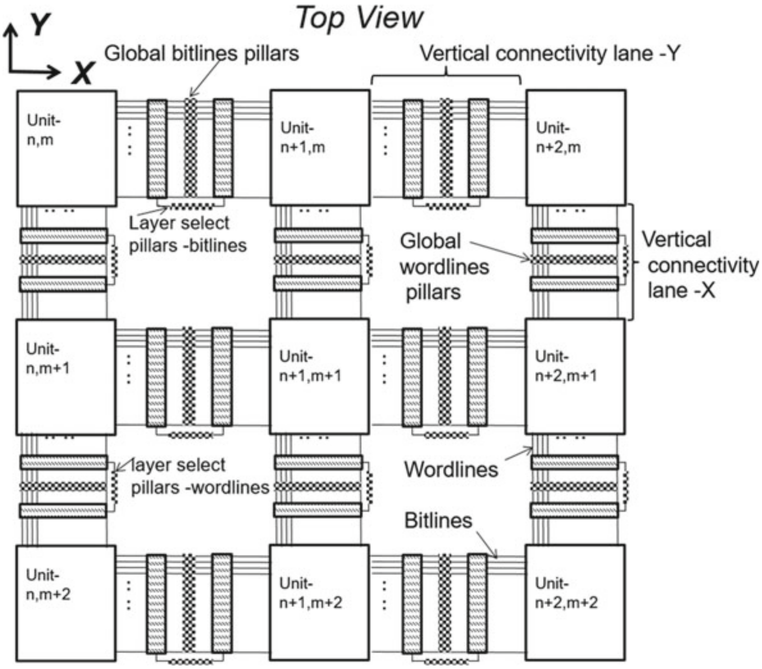
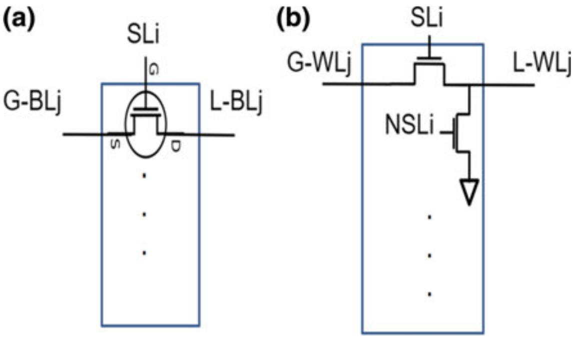


Fig. 15.9 Exemplary 3 × 3 units region of the bit-cell array stratum

bit-lines and the perpendicularly-oriented word-lines allow control of the individual bit-cells within a unit. These memory control lines extend across units, yet as part of a connectivity lane they have a connectivity control, called layer select, as illustrated in Fig. 15.10a, b. The local bit-line of line j (L-BLj) will be connected to the corresponding global bit-line j pillar (G-BLj) through a select transistor controlled by layer select i (SLi). The connection lane between units, carry the corresponding layer select per unit per control line, controlling the connection, to the global pillar of that control line (bit-lines, word-lines).

Fig. 15.10 **a** Bit-line layer select. **b** Word-line layer select



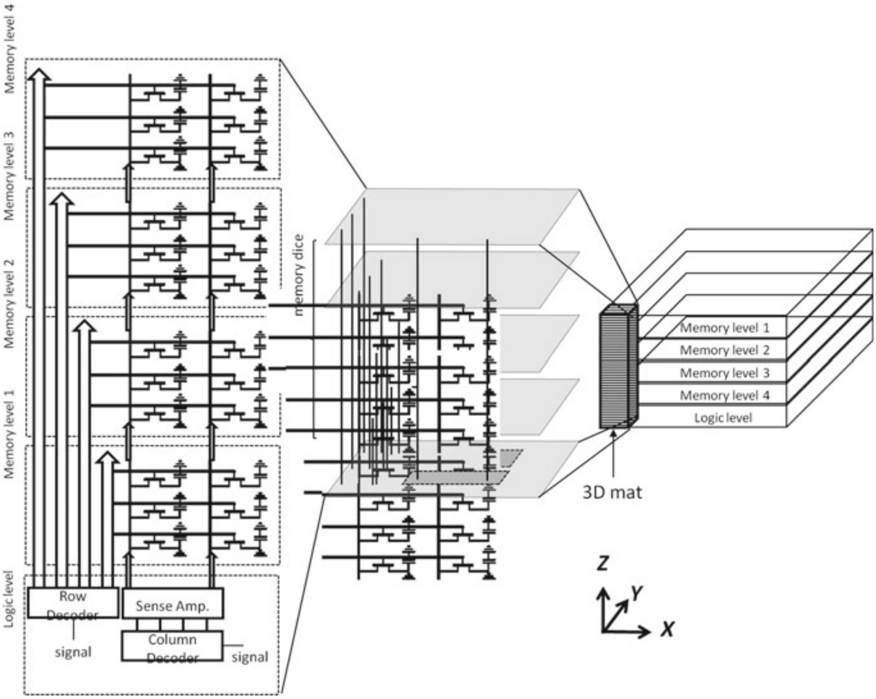


Fig. 15.14 3D illustration of memory strata formed by successive stacking

stack. Additionally, a parallel high speed, data transfer between strata in the stack can be facilitated using the proposed architecture.

The memory stack design also includes pass-through pillars, which allow transferring signals through it such as to allow synchronization of the memory control strata for the case in which one is under the memory strata and another is overlying it. The pass-through pillars could be used also for I/O when a processor stratum is placed underneath the memory strata as the base wafer, while the I/O stratum is placed at the top of the SoC stack. Thermal vias could be included to help thermal management.

Additional power delivery pillars can be included in the memory stack both for supplying memory power needs and to deliver power through the memory stack to strata underneath it.

An important advantage of this proposed architecture is the ability to form a per-unit redundancy. By having a redundancy stratum and proper circuitry in the memory controller, the layer select decoding circuit could include a mapping table to skip a ‘bad’ unit stratum and replace it with a unit in the redundancy stratum. Having thousands of units per die allows repair even in memory strata with tens of defects. This concept could also be used for field repair, providing a valuable advantage of this architecture.

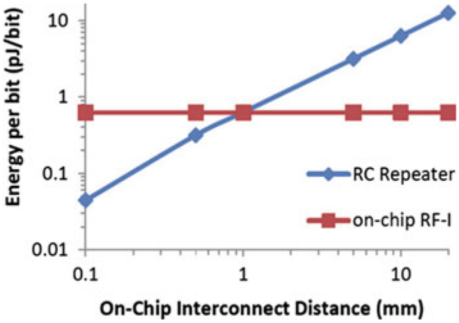
15.6 3D Heterogeneous Integration Enables Electromagnetic Waves Interconnects

A modular 3D IC system, as suggested here, that utilizes arrays of units each with its unit 3D memory cell block, memory control circuit block, processing logic block, and I/O block, needs good in-plane (X-Y) lateral interconnect with high throughput and low power consumption for system level functionality. While the out-of-plane (Z) vertical interconnects are formed having vertical vias with nano-meter and up to micron sizes and relatively short heights, the interconnect length in the horizontal in-plane direction (X-Y) remains at millimeter sizes, from die level (3–16 mm, for X and Y sides), reticle level (20–30 mm), to multi reticles, and up to wafer sizes (60–300 mm). Clearly the interconnect challenge is greater for the X-Y interconnect and the propagation delay and power dissipation using low-resistance metals such as copper and low-k dielectric material will end up impeding the 3D system performance.

As presented in Figs. 15.1–15.2b, today's interconnects are the limiting factor of computing electronics. The simple voltage representation of a logic signal is very sensitive to the interconnect RC. The most effective path to overcome this fundamental physical limitation is to shift from voltage logic representation to modulated electromagnetic (EM) wave of signal representation [15, 16] (Fig. 15.15).

The spectrum of the EM wave could be selected to fit the average target distance and the access to the appropriate technology. 3D heterogeneous structures could open the door to EM interconnects by adding strata of RF or Optical drivers, receivers, modulators and waveguides. In conventional 2D devices the cost of new nodes development and infrastructure drove vendors to focus their development to the most critical functions of logic and SRAM. Accordingly, any design targeting advanced manufacturing nodes must exclude anything other than what leading fabs include in their technology offering, which would be logic gates, SRAM and some I/O and basic support for analog function. The implication is that in advanced nodes RF or optical functions are not available and X-Y interconnects would be limited to RC Repeaters. Adapting 3D heterogeneous integration enables adding strata that could be built in other types of fab, such as RF-SOI lines, enabling the use of them

Fig. 15.15 RF-I will crossover the energy efficient curve of the RC repeater and become more energy efficient above a 1 mm interconnect distance at a 16 nm CMOS process [15, 16]



for the global X-Y interconnects. Within some technology parameters, the cross over from RF to Optical could be at over 30 cm [15, 16] (Fig. 15.16).

Wafer availability and cost could have a strong impact upon such choice. It is our assessment that the adoption of the 5G wireless communication standard and the increased use of wafers for RF applications would make RF-I the preferred choice for many applications. Figure 15.17 provides some benchmarks for these interconnect options [17].

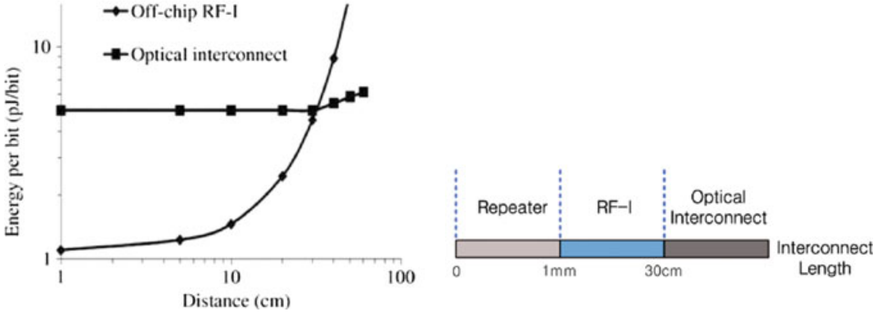
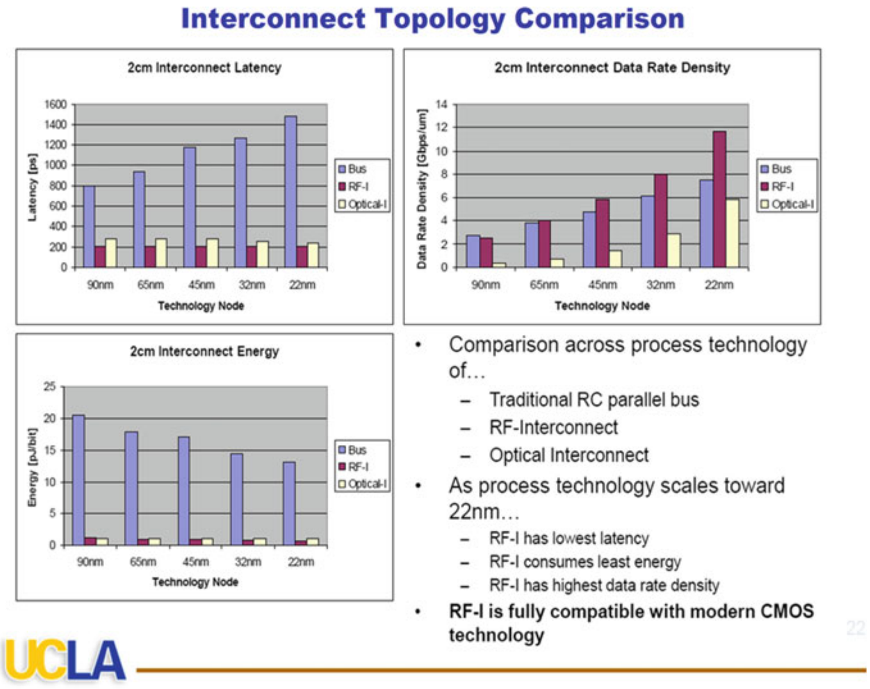


Fig. 15.16 RF-interconnect (RF-I) versus optical interconnect [16]



Interconnect Topology Comparison

- Comparison across process technology of…
  - Traditional RC parallel bus
  - RF-Interconnect
  - Optical Interconnect
- As process technology scales toward 22nm…
  - RF-I has lowest latency
  - RF-I consumes least energy
  - RF-I has highest data rate density
- **RF-I is fully compatible with modern CMOS technology**

Fig. 15.17 Benchmarks for 2 cm interconnects [17]

An important aspect of the monolithic 3D technologies as presented in Chap. 8 is the enablement of heterogeneous integration, in which one level (wafer) is produced using processes and materials to fabricate logic devices while another level (wafer) is produced using different processes and different materials to fabricate on-chip RF or optical interconnect devices. Furthermore, these levels (wafers) would likely be made in different fabs. Then, using a layer transfer process, one level is transferred over the other enabling fine vertical (3D) integration between the two.

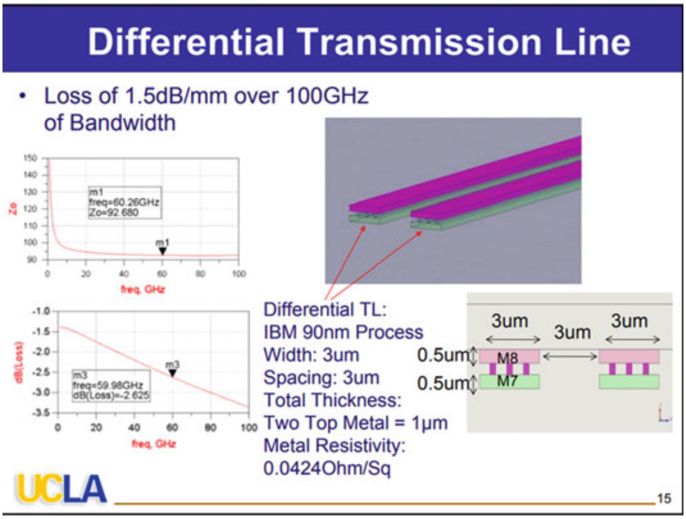
The on-chip RF or optical interconnect level could include more than one sub-level, for example, such as a passive photonic device level(s) for signal routing such as wave guides, photonic crystals, and resonators, and an active device level(s) such as a photo-detectors and a light sources (example e.g., lasers). The photo-detectors and light sources can each reside in its own different levels or they can be in the same level but with the two made with different substrates knitted together side by side. For example, the photodetector may be based on germanium, the light source may be based on a III–V semiconductor, and the passive devices may be based on silicon (core)-silica (cladding) structures.

15.7 Ultra Scale Integration (>1000 mm^2)

The key challenge of large reticle size or wafer level integration is yield. 3D integration may include multiple redundancy structures and repair techniques [13, 18–20] which could be used for robust RF and optical interconnected 3D system. Another alternative is to leverage the fact that RF transmission lines and optical interconnect waveguides are relatively large structures that have a very high yield with today process capabilities. The benchmarks of Fig. 15.17 were based on transmission lines having a 6 μm pitch, compared to advanced semiconductor process having less than 60 nm pitch. Optical waveguides use larger than a micron pitch lines as well. These large structures could be processed at very high yield while the drive electronics could be structured with redundancy for yield robustness (Fig. 15.18).

To allow ultra-scale integration of structures larger than a single reticle, the connectivity structure should extend over more than single reticle (>30 mm). Techniques to use optical lithography to pattern large areas greater than the full reticle field by ‘stitching’ multiple reticle patterns that had been projected independently are known in the industry, and are currently used for Interposer lithography and other applications. Alternatively some lithography tools are designed to support large area projections [21, 22].

Additionally, some prior work suggests integrating systems using an interposer with optical waveguides [23]. An additional alternative is to pre-test the RF or the optical interconnect components allowing the use of the concept of Known-Good-Die to wafer level die-to-wafer 3D integration by pretesting the RF or the optical interconnect fabric before transfer over to the 3D system. This could be efficiently implemented with the use of a generic RF or optical interconnect which could be



Differential Transmission Line

- Loss of 1.5dB/mm over 100GHz of Bandwidth

Differential TL:
IBM 90nm Process
Width: 3um
Spacing: 3um
Total Thickness:
Two Top Metal = 1μm
Metal Resistivity:
0.0424Ohm/Sq

Fig. 15.18 Transmission line example

produced in volume and pretested before use for the specific application. Another option is to avoid the physical interconnects and use wireless interconnects [24, 25].

The use of RF could include the use of differential signaling, which would help reduce the cross talk and interference effects, thus allowing lower supply voltages, and other advantages. The previous concepts for interconnection fabrics could be adapted to use differential transmission lines [26, 27].

Figure 15.19a, b illustrate a 3D system which include X-Y horizontal interconnection fabrics at relatively the upper level of the structure. In general, the horizontal interconnection fabric could be engineered in the middle level of the 3D system or at any other level. Placing it in the center could be advantageous in some systems by having a compute structure on both sides (under it and overlying it) thus allowing shorter vertical paths from the computing structures to the X-Y horizontal interconnection fabric. Figure 15.19a illustrates the structure as a generic continuous array of

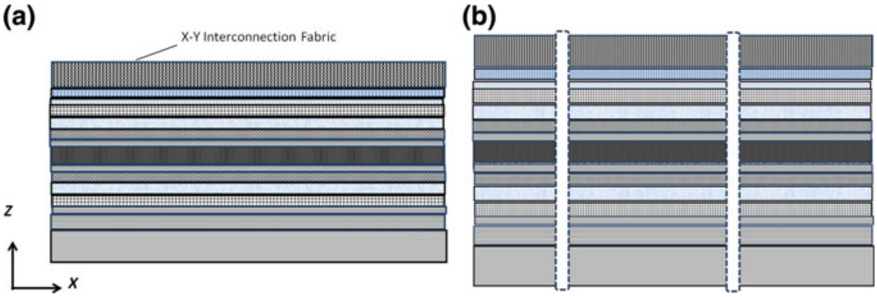


Fig. 15.19 **a** 3D heterogeneous integration. **b** 3D structure diced to smaller devices

cores, each with its own memories on top, and X-Y connectivity structure allowing data transfer between cores. Figure 15.19b illustrates the structure after being diced to smaller devices. There is a commercial value in building a generic computing platform to be produced in high volumes, which could be later used to specific market needs by dicing the generic structure according to the computing power needed for the target application.

A 3D system could include X-Y waveguides or transmission lines with configurable connectivity such as Single Write Multiple Read (SWMR), Multiple Write Single Read (MWSR), or even Multiple Write Multiple Read (MWMR). Connectivity fabrics where waveguides/transmission lines are designed for MWMR [28–30] simplify the configuration of its resources by adapting who gets to 'write' into a specific waveguide and who gets to read based on considerations such as yield and sizing (customization) (Figs. 15.20 and 15.21).

The concept of MWMR allows flexible use of the interconnection fabric in which compute units can sign in and sign out into the system's overall computing fabric.

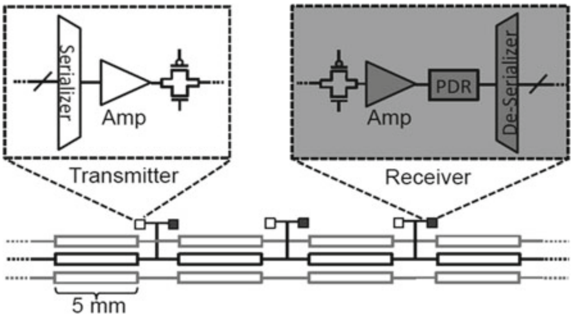


Fig. 15.20 RF interconnect with MWMR

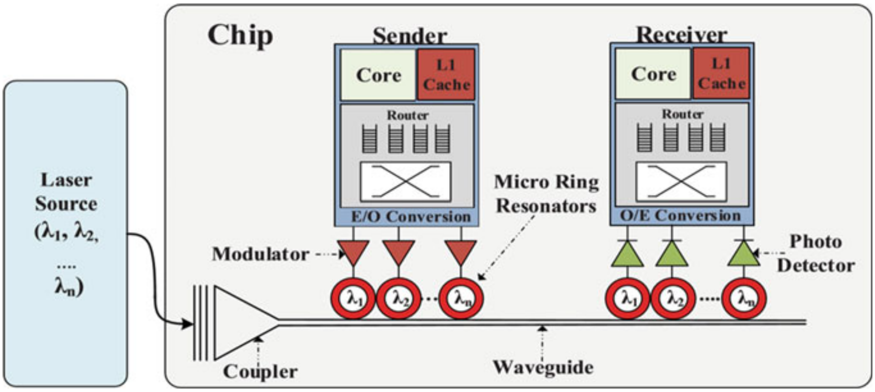


Fig. 15.21 Optical interconnect with MWMR

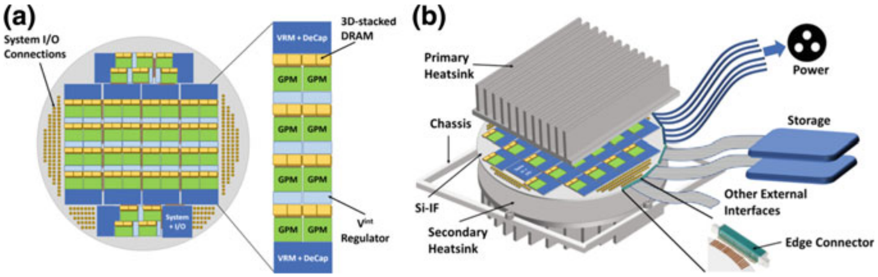


Fig. 15.22 **a** Waferscale GPU with 42 GPM unit (2 redundant). **b** Overall structure [34]

Such an architecture would be very tolerant to yield loss and to system reconfiguration based on yield or field customization.

The concept of wafer scale integration ("WSI") has been considered and explored over many years. It was never adopted due to the challenge of defects and due to the success of scaling. There is more interest these days as conventional scaling has slowed and with the growing interest in Artificial Intelligence (AI) and brain inspired architectures [31–33]. Recent work [34] demonstrated over 100× Energy Delay Product (EDP) for such wafer scale integration of GPU even without the use of EM interconnect. Figure 15.22a, b illustrate such wafer-scale demo.

The concept of leveraging 3D integration for wafer scale integration, or for multi reticles or multi die integration is extending the idea commonly used for memory repair. Memory repair utilizes the availability of redundant similar function memory cells designed with similar access time. Use of EM interconnect with arrays of computing units each with its own memory is similar. The functional units are equivalent and the X-Y EM connectivity is generally dominated by the delay converting a voltage to or from the EM signal, and is far less dependent on the location of the unit within the array. Accordingly redundancy would work well just as it is commonly used for memory repair (Fig. 15.23).

This enables wafer-scale integration and resolves the fundamental limit behind Moore's Law—yield.

It was yield that was driving the cost of integration up beyond some level of integration due to defect density. Once redundancy can be effectively used, defects do not limit the device size, allowing wafer-scale integration with an additional 1000× potential Energy-Delay product advantage.

15.8 Cooling

3D Systems such as those presented herein commonly generate heat while in operation, which must be managed to protect the system from heating up and affecting its operation. Figure 15.22b illustrates air cooling techniques for wafer scale system [34]. The next level of heat removal is the use of Microfluidic Cooling [35–37].

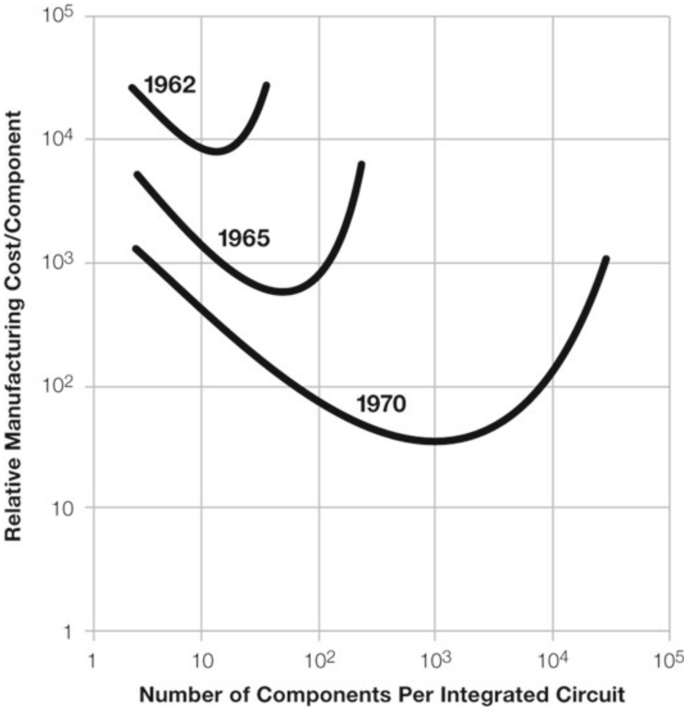


Fig. 15.23 The famous chart resulting in Moore's Law

MC has been proposed and is now used with some 3D devices at the device level (Fig. 15.24).

An additional advantage of the 3D wafer level heterogeneous integration of wafer-scale systems is the option to naturally form a micro-fluid cooling fabric in the substrate. Instead of forming micro-fluidic channels at the individual device level

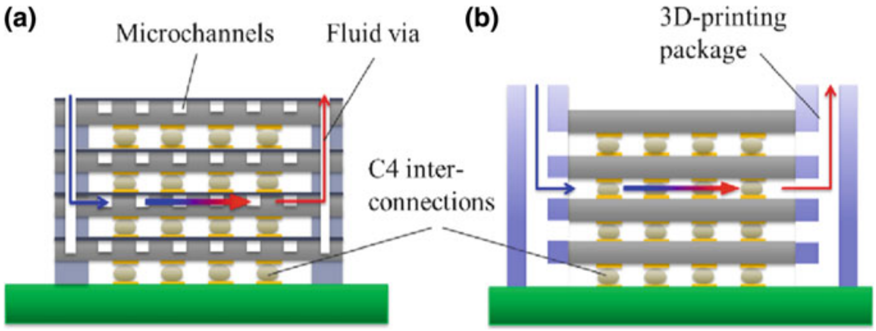


Fig. 15.24 **a** Diagrams of microchannel cooling, **b** fluidic chamber with 3D-printing package

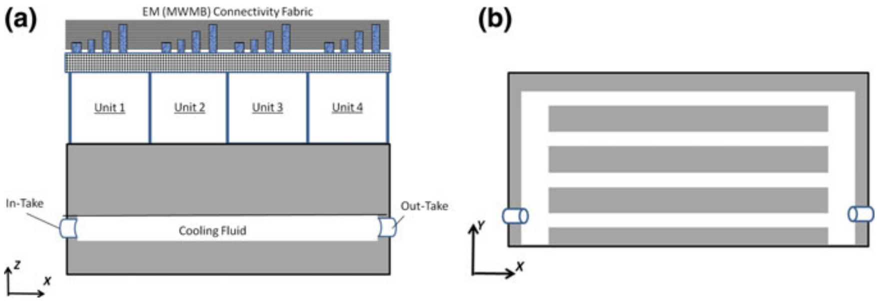


Fig. 15.25 **a** Wafer level microchannel cooling. **b** Horizontal cut view of (**a**)

and connecting them for the system level, the micro-fluidic cooling system could be formed at the wafer substrate and provide effective cooling system to the wafer scale system.

Figure 15.25a illustrates an X-Z cut view of such large scale 3D device integration with a substrate constructed to support fluid cooling. The illustration includes four computing units each with its own memories and connectivity (MWMR) to an EM connectivity fabric. The channeled silicon substrate could include micro-channels designed with fluid in-take and out-take. The substrate could be preprocessed to include the micro-channels at the wafer level, or bonded afterward to a micro-channel structure, for example with silicon to silicon bonding. Figure 15.25b is an X-Y cut-view through a micro-channel structure of the cooled 3D device. The micro-channels could be formed by etching trenches using conventional semiconductor processes into the micro-channel structure and then bonded to the wafer substrate. The micro-channel structure and a thinned wafer substrate could be slightly oxidized to enable a silicon dioxide to silicon dioxide bond if required by engineering and production constraints. Alternatively, the inner surface of the micro-channel may be further protected by silicon nitride or other desired film in order to protect the device from the cooling fluid. The wafer substrate could be thinned down by conventional techniques such as grinding and etch prior to the bonding. Thinning the substrate post device processing down to 50 μm is common in the industry.

15.9 Summary

3D heterogonous Integration with modern high-precision aligners allows the system designer to utilize wafers sourced from different fabs to form a 3D system. Using such integration technology allows constructing a computing system that could be many orders of magnitude better than today’s 2D PCB–based integration technology.

By integrating memory on top of the processor logic, the memory wall can be overcome, resulting in a 1000× better computing unit.