3-D Integration and the Limits of Silicon Computation

Dinesh Pamunuwa*, Matthew Grange*, Roshan Weerasekera* and Axel Jantsch+

*Centre for Microsystems Engineering, Faculty of Applied Sciences,
Lancaster University, Lancaster LA1 4YR, United Kingdom.

{d.pamunuwa,m.grange,r.weerasekera}@lancaster.ac.uk

+Dept. of Electronic, Communication, and Software Systems,
Royal Institute of Technology (KTH), Forum 120,SE-164 40 Kista,Sweden

axel@kth.se

ABSTRACT

The *intrinsic computational efficiency* (*ICE*) of silicon defines the upper limit of the amount of computation within a given technology and power envelope. The *effective computational efficiency* (*ECE*) and the *effective computational density* (*ECD*) of silicon, by taking computation, memory and communication into account, offer a more realistic upper bound for computation of a given technology. Among other factors, they consider how distributed the memory is, how much area is occupied by computation, memory and interconnect, and the geometric properties of 3-D stacked technology with through silicon vias (TSV) as vertical links. We use the *ECE* and *ECD* to study the limits of performance under different memory distribution, power, thermal and cost constraints for various 2-D and 3-D topologies, in current and future technology nodes.

1. INTRODUCTION

The traditional scaling methodology towards faster processors and higher frequency has been hampered by unfavorable interconnect scaling characteristics, increased sub-threshold gate leakage, higher power densities and rising costs. As a consequence the increase of transistor count as described by Moore's law is translated into highly parallel, low-complexity cores supported by high-throughput packet switching Networks-on-Chip (NoC).

Due to their compact geometry, 3-D integrated systems hold promises to significantly reduce latency, power consumption and area, while increasing bandwidth. In the following we quantify the potential and limits of 3-D integration by analyzing the theoretical performance of various 2-D and 3-D topologies. Figure 1 shows how the geometric distance between cores grows very differently in 2-D and 3-D structures with the number of cores. Since for global and long distance communication the geometric distance translates linearly to latency, we can expect to cut communication latency by 50%. A number of recent studies of communication performance in 3-D structures [1, 2, 3] demonstrate the significant potential of 3-D integration technology for reducing power consumption and increasing performance.

3-D integration enables stacking of memory on top of processors, thus realizing a direct low latency and high bandwidth memory access link. However, to exploit the benefits, the memory architecture has to be adapted to allow for multi-port, parallel memory access. Several recent studies have explored various memory and cache architectures while exploiting the third dimension. For instance F. Li et al. [4] propose a 3-D distributed L2 cache and observe a 50% access latency reduction, essentially due to shorter wires within the L2 cache. G. Loh [5] explores the effect of parallel memory access by means of multiple memory controllers and ranks in a 3-D stacked DRAM based memory architecture and reports a performance increase of more than 280% over a conventional memory architecture for a set of benchmark applications.

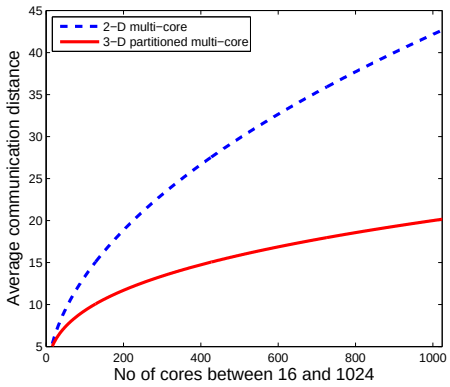

Figure 1: The average geometric distance for a multi-core system for a 2-D and a 3-D realization.

We adopt T. Claasen's notion of *intrinsic computational efficiency* of silicon [6]. The intrinsic computational efficiency is obtained when all the silicon area is filled with elementary operations, say 32-bit adders, and no area is "wasted" for data communication and control. Figure 2 shows how the intrinsic computational efficiency, measured in millions of operations per second per Watt, is increasing with technological progress. The left part of the curve is copied from T. Claasen's original paper [6], while the right part is based on our own model, as introduced below. For comparison we have marked the performance of two recent multi-core processors from Tilera Inc. [7].

Different architectures such as micro-processors, DSPs, FPGAs and custom hardware, will approximate this line to a higher or lower degree depending on how well an application matches the architecture and how much flexibility is built-in. But no real processing unit can match or exceed it. Larger and more general purpose processors exhibit a greater gap because they utilize more area and power on interconnect, control and provision of programmability.

In our model we assume that in a 3-D topology, DRAM is used as embedded memory because it can be placed on a separate die,

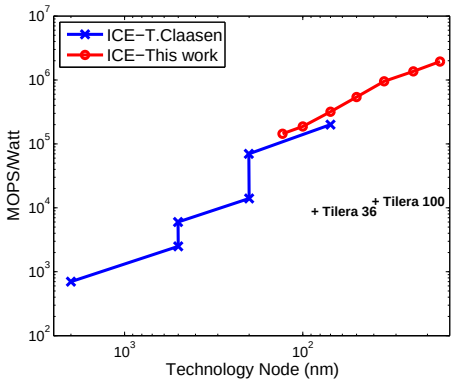

Figure 2: The Intrinsic Computational Efficiency of Silicon, where the the entire die area is filled with 32-bit adders. The difference between T. Claasen's projection and ours lies in the architecture of the 32-bit adder, leading to a minor difference in the energy/operation

thus leveraging on the capability of 3-D to integrate different process technology in the same system. In section 2, we introduce and motivate our analytical model for performance comparison. Then, we describe the technological parameters for performance, power consumption and area for the 2-D and 3-D topologies (section 2.3). Also, we explain our scaling methodology. Section 3 delves into the impact of how memory is distributed in the system. Not surprisingly a distributed memory exhibits higher performance than a centralized memory. However, after distributing 80% of the memory, further diffusion of the storage has little added benefit. In section 3.1 we become more concrete by assuming specific system sizes, frequencies and thermal/power budgets. This allows for analyzing performance limits under realistic physical constraints. We follow by discussing cost in section 3.2. Finally, we provide further discussions of the model and our conclusions in section 4.

2. MODELING SILICON EFFICIENCY

Intrinsic Computational Efficiency (ICE) of silicon [6] is the number of 32-bit add operations per Joule, or the number of operations per second per Watt. The *ICE* reflects the amount of computation that can be done within an energy envelope, but it does not measure the amount of computations per area or per volume. We now define the *Intrinsic Computational Density (ICD)* as the number of 32-bit adders that fit into one mm^2. Figure 2 shows the intrinsic computational efficiency as a function of technology nodes. Theo Claassen's plot from 1999 is repeated and for comparison the *ICE* figures of our model (section 2.3) for technology nodes between 180 nm and 16 nm are added.

2.1 Adding Memory and Communication

ICE and *ICD* give the upper bound of what amount of computation can be done with a given silicon technology under the assumption that the entire area is densely packed with computation units. However, we also need to account for memory, where the data are stored before and after processing, and for interconnect, which allows the data to move between processing units and memory. In the following we study variants of *ICE* and *ICD* under less ideal assumptions in different 2-D and 3-D configurations. In particular we explore the following factors and configurations.

To include the effects of memory and interconnect, we define the *Effective Energy (EE)* for a 32-bit addition as follows.

$$EE^{\text{tn}}_{\text{arch}} = E^{\text{tn}}_{32} + \mu_T\big(\omega(e_1 + \Delta \times \text{E\_int}^{\text{tn}}_{\text{arch}}) + (1\text{-}\omega)(e_1 + \text{E\_int}^{\text{tn}}_{\text{arch}} + \text{E}_{\text{offchip}})\big) \quad (1)$$

for a given technology node, tn, and a given architecture, $arch$, $\in \{2D, 3D2, 3D4, 3D8, 3D16\}$. The three main terms correspond to the energy consumption of an addition, of on-chip memory access and of off-chip memory access, respectively.

- e_1 is the amount of energy it takes to read or write one 32-bit word in on-chip SRAM.
- $\text{E\_int}^{\text{tn}}_{\text{arch}}$ is the energy it takes to transport one 32-bit word from a non-adjacent on-chip memory to the local cache. For example, if the total silicon area is 400 mm^2, we have

$$\text{E\_int}^{\text{tn}}_{\text{arch}} = \begin{cases} (10+10)e_2(tn) & \text{if arch}= 2D \\ (7.07+7.07)e_2(tn) + e_3() & \text{if arch} = 3D2 \\ (5+5)e_2(tn) + 2e_3() & \text{if arch} = 3D4 \\ (3.5+3.5)e_2(tn) + 4e_3() & \text{if arch} = 3D8 \\ (2.5+2.5)e_2(tn) + 8e_3() & \text{if arch} = 3D16 \end{cases}$$

- e_2 is the energy it takes for a 32-bit word to be transported 1 mm horizontally in a given technology.
- e_3 is the energy it takes to move a 32-bit word from one vertical level to the next via a set of TSVs.
- $\text{E}_{\text{offchip}}$ is the energy it takes to get off-chip and to read or write the off-chip memory. It includes the I/O drivers, the inter-chip communication and the energy consumption of the memory chip.

The idea of E_int is to capture the communication energy in different architectures to get from an arbitrary point in the system to a particular point at the system boundary. For a 2-D 20×20 mm^2 die, the distance is on average 10 mm in each dimension, hence it is 20 mm. For a 3-D structure we have to traverse half of the vertical levels on average. E.g. for a 3D2 we have to traverse 2 vertical levels.

Thus, the effective energy *EE* gives the required energy for a 32-bit addition if memory access and communication is taken into account. The factors μ_T, ω and Δ are abstractions of architectural choices and features. Based on EE we define the *Effective Computational Efficiency (ECE)* as $ECE^{tn}_{arch} = \frac{1}{EE^{tn}_{arch}}$ which gives the amount of computation we can do within the energy envelope of 1 Joule; or the amount of computations per second we can do within the power envelope of 1 Watt.

- **Ratio of computation to memory μ:** We distinguish between the temporal ratio μ_T and the spatial ratio μ_S. The relative number of memory accesses for each operation is μ_T, while μ_S is the number of memory words in the system for each operator. With memory we mean SRAM, caches and the like but also off-chip DRAM. If $\mu_T = 1$, for each operation there is 1 memory access. Typical values will be between 1 and 3. On the other hand, the amount of memory is usually much higher than the operators. Hence, typical values for μ_S are between 1000 and 10000 as we discuss later.
- **Ratio of on-chip versus off-chip memory ω**: If $\omega = 1$, all memory is on-chip; if $\omega = 0$, all memory is off-chip. In a 3-D topology, with on-chip we mean all dies in the 3-D stack.
- **Memory distribution factor Δ:**
  - $\Delta = 0$: completely distributed memory. The distance between a computation unit and the memory is always 0;

- $\Delta = 1$: completely central memory where the distance between a computation unit and the memory is always the diameter of the system (or off-chip)
- e.g. $\Delta = 0.05$: models a cache system where 95% of all memory accesses are local and 5% are far away.

This parameter models the communication required to write to and read from memory. If the memory is completely distributed, we assume all memory reads and writes are local and no long-range communication is required. Obviously, this is a simplification but any specific architecture-application pair can be characterized by a Δ value between 0 and 1, denoting the amount of global on-chip communication occurring.

- We explore different **2-D and 3-D topologies** but we typically compare systems with the same total silicon area. E.g. if the total area is 400 mm^2, the configurations considered are 2D: one plain silicon die of size 20×20 mm^2; 3D2: 2 stacked dies, each 200 mm^2; 3D4: 4 dies of 100 mm^2; 3D8: 8 dies of 50 mm^2; 3D16: 16 dies each 25 mm^2.

2.2 Effective Computational Density

Similarly, the area cannot be filled with computational units only. We need to take memory and interconnect into account as well. We define the *Effective Area (*EA*)* as follows.

$$EA^{\text{tn}}_{\text{arch}} = A^{\text{tn}}_{32} + \mu_S\ \omega\ a_1\ + \sigma\ \text{A\_int}^{\text{tn}}_{\text{arch}} \qquad (2)$$

EA is defined similarly to *EE* but the off-chip component is omitted since we do not include the area for off-chip memory. Again, with "off-chip" we really mean "out-of-package". Different dies in a 3-D stack are considered "on-chip".

- a_1 is the area for a 32-bit memory word. Depending on the geometry we assume either SRAM or DRAM memory. For a 2-D system we use the area of embedded SRAM, while for a 3-D system we use DRAM. Concretely we use $60F^2$ area for one SRAM cell [8] and between $8F^2$ and $4F^2$ for DRAM cells [9], where F is the minimum feature size.
- A_int$^{\text{tn}}_{\text{arch}}$ is the interconnect area required for transporting a 32-bit word to memory.
- σ is the interconnect sharing factor. If $\sigma = 1$, no sharing takes place and every operator has its own, private interconnect across the system. If $\sigma = 0$, the interconnect is optimally shared and the interconnect area per operator is 0. In analogy to μ_S it gives the ratio of area occupied by operators versus interconnect. Typical values are between 0.01 and 0.1. For instance the Tilera TILE64 [10] with 64 cores has 8 32-bit operators per core[1] For 512 operators on the 64 core Tilera die with an 8×8 mesh interconnect, the sharing factor is $\sigma = 16/512 = 0.031$.

  Note, that only the global interconnect for the dataflow is counted, while the local interconnect and global control lines are ignored.
- a_2 is the required area for a 1 mm long 32-bit bus.
- a_3 is the required area for 32 TSVs to connect from one horizontal layer to the next.

[1]Again, this is a simplification, because there are fewer operators but they are pipelined. In effect, 8 operations can be completed per cycle in the best case, motivating the number 8 that we use in this example.

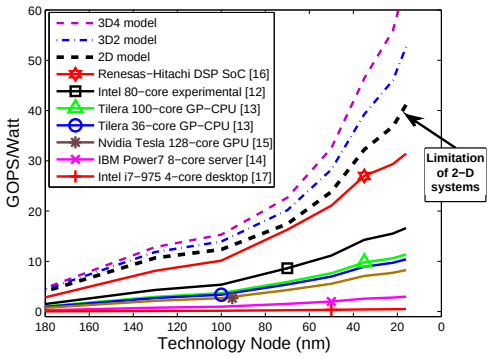

Figure 3: The Effective Computational Efficiency of recent multi-core processors [11, 10, 12, 13, 14, 15] (where markers indicate actual performance data) compared to our model.

For instance, if one 32-bit adder has the size 437.5 μm^2 in 50 nm technology and $\mu_S = \sigma = 0$, we get an $ECD^{50} = ICD^{50} = 2286.8 \frac{\text{operations}}{mm^2}$ and we can fill a 400 mm^2 chip with 914,720 adders. In a more realistic scenario with $\mu_S = 4000$ and $\sigma = 0.031$ we get 345 operators on a 400 mm^2 area.

2.3 Scaling circuits, devices and interconnects

The underlying parameters used in our models for energy and area efficiency are based on realistic circuit-level parameters extracted from published data, SPICE simulations, parasitic extraction tools, and consistent scaling methodologies [16]. We model global planar 2-D wires, TSVs, logical operations, memory transactions, thermal properties, transistors and leakage power.

2.3.1 Interconnect

We scale the global wires in silicon-based ICs for technology nodes 180 nm down to 16 nm using a similar methodology to the authors of [17] where wire parasitics, including parallel plate, fringe and coupling terms, and resistance are extracted from field solver simulations and compact models fitted to extract parameters for future technology nodes given the global wire dimensions, barrier thickness, spacing, resistivity of the medium, vertical and horizontal dielectric constants (including low-k and high-k) and the switching probability of the surrounding wires. Using the *RC* characteristics of the wires, typical repeater insertion strategies, and scaling supply voltages, we determine the energy-per-bit for die area dependent wire lengths across technology nodes from 180 nm down to 16 nm.

For the purpose of extracting parasitics for the 3-D interconnect we simulate copper TSVs with a uniform circular cross-section and an annular dielectric barrier of SiO_2 or Si_3N_4 surrounding the Cu cylinder with a thickness of 0.2 μm. To be model the maturity of TSVs as technology progresses, we simulate radii of 10, 8, 6, 4, 2 and 1 μm and a constant length of 50 μm. The pitch of the TSVs is twice the radius to match planar global wire spacing trends.

2.3.2 Logical Operation and DRAM Scaling

The logical on-chip operations such as a 32-bit addition or SRAM read are modeled by using published data [18, 19] for a particular technology node and scaling the energy and area for future or past generations. The off-chip DRAM transaction energy is not a simple function of the feature size, and depends on the DRAM architecture, its peripheral circuitry and also characteristics of off-chip

drivers, terminations, and chip, package and board trace parasitics. We have used the Micron System Power Calculator [9] to estimate the average off-chip read/write power for different generations of DRAM, from SDRAM to DDR3. We have matched the DRAM generation to the technology node, such that 180 nm corresponds to SDRAM and 16 nm to DDR3.

2.3.3 Power, Leakage and Thermal Models

We have created compact thermal models based on material dimensions, conductivities and cooling strategies to provide a power-related temperature analysis in our 2-D and 3-D architectures. We extracted leakage power trends across technology nodes using SPICE simulations with Berkeley Predictive Technology Models [20] for bulk CMOS transistors for a temperature range of 0^oC to 200^oC to develop temperature-dependent current sources for operational die power in our analysis. We mainly use forced convection air cooled heatsinks at the top of the package with a typical ball-grid array package, but we also consider the added heat removal benefit of microchannel liquid cooling between die layers.

3. DESIGN SPACE EXPLORATION

Varying the parameters (further described in section 2) in our model such as the amount of on-chip cache versus off-chip memory transactions and memory distribution factor (effectively how far away is the cache resource to the computational unit), can allow for virtually any processor architecture to be represented. Figure 4 shows the *ECE* for various topologies when Δ, representing the proportion of centralized memory, varies between 0 and 1. For all topologies a centralized memory drags down *ECE* significantly from over 60 GOPS/W to about 5-10 GOPS/W. Hence, there is a benefit from distributing memory, but only a distribution of $\Delta < 0.2$ has a significant effect. This benefit from distribution is more pronounced for a 2-D topology. Going from Δ=1 to Δ=0.1 improves *ECE* for 3D16 by a factor 5, while the improvement is 8 for 2D. Intuitively the reason for this is that the energy of transporting data across the chip to a central memory is much lower for a 3-D topology. Hence, if it is difficult to decentralize most memory accesses, the penalty will be lower for 3-D. However, the impact of centralized memory on performance becomes steeper for more advanced technologies. The effect is apparent for a 3D16 topology. While the difference in performance between Δ=0 and Δ=1 is a factor 5.4 for 180 nm technology, it grows to a factor of 34 for a 16 nm technology. Hence, even if a 3-D topology can mitigate the cost of centralized memory, it is still growing exceedingly as

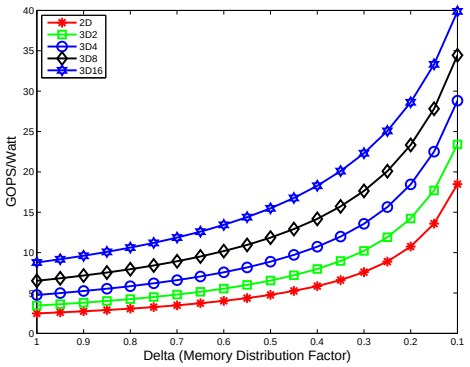

Figure 4: The effect of on-chip memory locality on the *ECE*

technology advances due to the inverse effect on the performance of logic versus interconnect as a result of scaling.

For the purpose of a comparative study between 2-D and 3-D topologies, we have distributed a single processor over 3-D layers of 2, 4, 8 and 16, so each 3-D die in a 16-layer processor will have $1/16^{th}$ the power of the total 2-D processor. To compare architectures, we consider concrete system configurations under power, frequency, area, and thermal constraints. Our scenarios mainly model cache systems where 95% of the memory transactions are on-chip (ω=0.95) and local (Δ=0.05). Furthermore, to represent a realistic system we require a certain amount of on-chip memory, in this case we set our μ_S parameter to 2000 words per operator, similar to the Tilera TILE64 [10] processor. Adjusting the amount of cache in either direction will directly affect the number of operators that can be squeezed into the die. We assume an interconnect sharing ratio, σ, similar to the interconnect area of a large mesh-based NoC or many-core processor such as the TILE64.

3.1 Performance

Our model exposes the potential of 3-D stacked systems, which mainly stems from (1) the possibility to integrate dense DRAM tightly into the multi-core architecture, and (2) from the more power efficient interconnection in the third dimension, which essentially is due to shorter geometric distances. Theoretically a 3D16 topology offers 2.4 times higher performance per Watt than a 2-D topology and for every doubling of the stack height, we see a 20 to 30% increase of the performance per Watt figure. This relationship is encapsualted in Figure 5(a), where the maximum throughput is shown for increasing power constraints in a 400 mm^2 16 nm processor. However, when cooling limitations are considered, we find that 3-D ICs above eight layers can only dissipate a maximum of 15 W with conventional air-cooled heatsinks, where stacks of up to four layers can consume up to 50 W of total power. In most realistic cases, the operational power of a device will be dictated by the technology, packaging, environment and the application rather than their absolute maximum limitations.

To understand realistic performance limitations of 2-D and 3-D architectures within a given domain, we have constrained each topology to operate below an absolute maximum temperature of 100^oC at any point in the structure for an upper power limit of 100 W. 3-D topologies are constrained by their thermal performance more so than 2-D systems operating under the same power budget. Figure 5(b) plots the maximum throughput versus the number of die layers given the maximum thermal ceiling for each topology. Further inter-die cooling strategies such as fins, interposers and microchannel cooling have been shown to reduce temperatures of air cooled systems by up to 30 % and it is likely, as the authors of [21] have also concluded, that additional cooling, such as liquid microchannels between die layers will be required for high performance logic-on-logic die stacks. Therefore, in Figure 5(b) we have shown the effect of different cooling strategies as a percentage improvement over air-cooled heatsinks to depict what may be achievable in the future with 3-D systems. The optimal topology for throughput at the maximum thermal design power (TDP) mainly lies at a 2-layer 3-D system and when additional cooling is considered the apex intuitively shifts to larger 3-D stacks as the maximum power-per-die improves.

It is clear that large 3-D systems operating at their thermal ceiling are sharply limited by lower power constraints necessary to maintain the thermal integrity of the package and logic. Leakage contribution to the total power consumption increases as feature size reduces and is more prevalent in 3-D topologies due to higher temperatures, which in turn degrades the the theoretical maximum perfor-

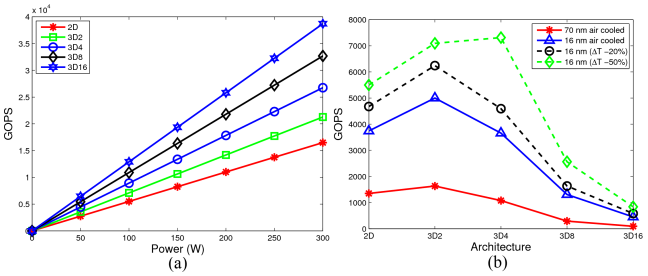

Figure 5: (a) The maximum throughput for our topologies constrained by power and (b) The maximum throughput with a thermal ceiling of 100^{o}C and 100 W. Dashed lines show additional heat removal beyond conventional air cooling

mance in larger stacks. However, the thermal junction-to-ambient package resistance in a two-layer 3-D system is still low enough that a two-layer 3-D system can attain higher throughput than an equivalent 2-D system. Furthermore, we find that a four-layer 3-D computational system at 35 nm has a performance advantage of 16% over the same system instantiated in a 2-D package two technology generations lower at 16 nm. This means that a processor design in 3-D with smaller numbers of layers can achieve equal or higher performance without significant investment in further technology node shifts. Further, Figure 5(b) shows the performance given the maximum power for each architecture, when in fact the design of a processor may typically depend on the application requirements, falling below its maximum TDP. Low-power mobile processing is one such application that seems especially suited to 3-D ICs due to stringent power, area, and performance requirements. Under low-power constraints, larger 3-D systems of up to 16 layers will fall within their thermal budgets and can provide 2-3 times the performance per Watt of a similar 2-D system.

3.2 Cost

3-D integration incurs significant costs related to stacking that are over and above the cost of a pure 2-D implementation, including the cost of the TSVs, test, pick and place of Known Good Dies (KGD) and bonding, and any additional cooling. Offset against this is the fact that the individual dies occupying the different layers are a fraction of the area that would be occupied by a single 2-D die, resulting in more dies per wafer as well as increased yield due to the smaller die area, with a significant drop in cost (per die in the 3-D stack), generally acknowledged to drop off as the 4^{th} power of die area [16]. The other main issue is that a significant (fixed) investment is required in shifting technology nodes, related to infrastructure as well as design. This investment is usually amortized over many production runs for large-volume processors and the cost-per-unit asymptotically approaches the variable cost (costs that are proportional to the volume of a given product) with increasing volume. As we have shown though, the performance gain achievable by reducing feature size with its accompanying costs can be matched or bettered by a 3-D implementation without a tech shift. Our focus in this section therefore is to answer primarily two questions: first, as the complexity of the system increases and the total silicon area grows, is there a point at which a system implemented in 3-D becomes more cost effective than a 2-D implementation, and if so, what is that area, and the corresponding system architecture? Second, is there a volume of units sold at which the total unit cost of a 2-D system including the design related Non-Recurring Engineering (NRE) cost of moving to that node (costs that cannot be billed directly to a single product) becomes equal to the total unit of an *equivalent* 3-D system implemented in the older technology, which does not have the NRE cost associated with feature size reduction. We call this volume the *cost-equilibrium volume*.

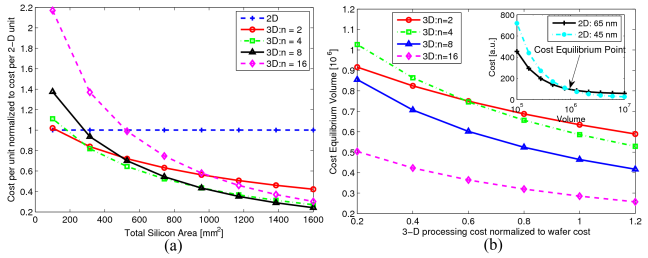

Figure 6: (a) D2W Variable cost (b) Variation of cost-equilibrium volume with 3-D process cost for an NRE design cost increment of 75% in a technology shift

In carrying out a comparative cost analysis we divide the variable cost into a die cost that includes material, labor, and process costs and a test cost. The die cost is a function of the wafer cost, number of dies per wafer and die yield. The yield has material-defected related (Y_m), systematic (Y_s) and random (Y_r) components which are complex functions of die area, and process related parameters including defect density and other statistically estimated quantities. We use typical values for MPU product families as reported in the ITRS [22] for these yields. The stacking cost for 3-D systems is estimated by factoring in the required extra mask layers and associated processing costs, the cost of pick-and-place and bonding as a fraction of the wafer cost. The test associated with selecting KGDs to bond onto the base wafer, and the yield drop due to stacking is also considered. Some of the model parameters can be sourced from the open literature [3, 22] while the main imponderable is the 3-D stacking cost as a fraction of the wafer cost. Based on information gathered from our involvement in 3-D integration projects, we have carried out investigations for a technology that can be approximated by a stacking cost that is 20% of the wafer cost.

The first question we posed is answered in Figure 6(a), which shows the costs of 3-D systems implemented using Die-to-Wafer (D2W) stacking normalized to the 2-D system cost for that particular silicon area. This normalized view clearly shows the cost effectiveness of 3-D vs 2-D; for smaller areas, 3-D integration is more expensive than a 2-D implementation and the greater the number of layers in the stack, the higher the cost. However, as the total silicon area increases, having more 3-D layers lowers the unit cost. That is because the cost increases approximately as the fourth power of die area, and for large areas a very low yield in a pure 2-D implementation can be contrasted with the much higher yield of the smaller individual dies in the stack, which more than compensates for the extra 3-D bonding cost and reduced yield in the stacking. For yield parameters provided in [22], the cost-equilibrium point for D2W cost is approximately 170 mm^2. This cost-equilibrium point changes with the 3-D stacking cost as well as defect density; the higher the defect density, the smaller the cost-equilibrium silicon area. It should be noted that the defect density of 0.13 per cm^2 used in this study is on the low side for cutting-edge technologies, and can be quite a bit higher, in which case 3-D would be even more attractive from a cost point of view.

To answer the second question, we estimated the design related portion of the NRE cost as being 75% higher when moving from 65 nm to 45 nm [23], which results in Figure 6(b). The inset shows the cost-equilibrium point for 2-D systems as being approximately

a million units. For the yield parameters used, the variable cost for implementing a 20 mm×20 mm system in 45 nm technology is about 24% of the cost in 65 nm technology, whereas it is 80% for implementing the system in a 2-layer 3-D configuration under the same 65 nm technology. The main graph shows what the cost-equilibrium point is for various 3-D implementations for a range of 3-D stacking costs and shows for example that even for a 3-D stacking cost of up to 40% of the wafer cost, approximately 900k units must be sold before the 45 nm 2-D implementation becomes more cost-effective than a 65 nm 4-layer stack.

4. CONCLUSIONS

We have developed the concepts of effective computational efficiency (*ECE*) and effective computational density (*ECD*) to study the limits of performance of 2-D and 3-D topologies with technology down to 16 nm. Our model provides an abstraction of real systems in order to provide an upper bound on the performance. As such, we have not considered the control structure including logic, local interconnect and registers (which is less significant in comparison with global communication). The lower the overhead of the control structure, the closer the performance of a real system to our predicted upper bound, which is encapsulated by the DSP [14] in Figure 3.

Another limitation is our focus on throughput as the main performance characteristic, while ignoring latency. Latency is much harder to capture at an abstract level since it is influenced strongly by many details of the architecture, arbitration policies and resource management strategies. In real systems the theoretical limits of throughput are often not achieved because raw capacity is overprovided and a lot of control logic is used to keep critical latency figures low. It can be noted however, that a main benefit of 3-D topologies is the lower latency of memory transactions since high capacity memory can be located much closer to the computation units. This may mean that 3-D systems come closer to their intrinsic performance limits than 2-D topologies.

In summary, although our model constitutes an idealization of systems, it still expresses correct trends and bounds of real systems and we draw the following main conclusions from our study:

- 3-D systems can attain 2 to 3 times higher *ECE* due to lower interconnect power;
- 3-D systems have one order of magnitude higher memory density due to DRAM integration which means they can accommodate more computation units in a given area with the same amount of memory;
- This allows for much higher performance but causes also very high power density. Die stacks of over four layers will mainly be suitable in low-power mobile applications or high-density memory stacks such as Flash memory and a controller.
- The same performance with the same power can be realized in 3-D topologies with much smaller area and at lower frequency.
- The added expense and yield loss associated with 3-D stacking can be compensated by higher individual die yields and reduced NRE investments allowing 3-D systems above 170 mm^2 to reach their cost-equilibrium point earlier than a 2-D system.

5. REFERENCES

[1] William J. Dally. Performance analysis of k-ary n-cube interconnection networks. *IEEE Transactions on Computers*, 39(6):775–785, June 1990.

[2] A. Y. Weldezion et. al. Scalability of network-on-chip communication architecture for 3-D meshes. In *Proc. Int. Symp. Networks-on-Chip (NOCS)*, pages 114–123, 2009.

[3] R. Weerasekera et. al. Two-dimensional and three-dimensional integration of heterogeneous electronic systems under cost, performance, and technological constraints. *IEEE Trans. Comp.-Aided Design of Integrated Circuits and Systems*, 28(8):1237–1250, 2009.

[4] Feihui Li, Chrysostomos Nicopoulos, Thomas Richardson, Yuan Xie, Vijaykrishnan Narayanan, and Mahmut Kandemir. Design and management of 3 D chip multiprocessors using network-in-memory. *ACM SIGARCH Computer Architecture News*, 34(2):130–141, 2006.

[5] Gabriel Loh. 3D-stacked memory architectures for multi-core processors. In *Proceedings for the 35th ACM/IEEE International Symposium on Computer Architecture (ISCA)*, June 2008.

[6] T. Claasen. High speed: not the only way to exploit the intrinsic computational power of silicon. In *Proc. Int. Solid State Circuits Conf. (ISSCC)*, pages 22–25, 1999.

[7] Tilera Corporation. Tilera home page. http://www.tilera.com.

[8] Stefan Lai and T. Lowrey. OUM - a 180 nm nonvolatile memory cell element technology for stand alone and embedded applications. In *Proc. Int. Electron Devices Meeting (IEDM)*, 2001.

[9] Jeff Janzen. The micron system-power calculator, 2009.

[10] S. Bell et. al. Tile64 - processor: A 64-core soc with mesh interconnect. In *Proc. Int. Solid State Circuits Conf. (ISSCC)*, pages 88–598, 2008.

[11] S.R. Vangal et. al. An 80-tile sub-100-W TERAFLOPS processor in 65-nm CMOS. *IEEE J. of Solid-State Circuits*, 43(1):29–41, 2008.

[12] D. Wendel et. al. The implementation of POWER7TM: A highly parallel and scalable multi-core high-end server processor. In *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pages 102–103, 2010.

[13] E. Lindholm et. al. Nvidia tesla: A unified graphics and computing architecture. *IEEE Micro*, 28(2):39–55, 2008.

[14] Y. Yuyama et. al. A 45nm 37.3GOPS/W heterogeneous multi-core SoC. In *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pages 100–101, 2010.

[15] Intel i7 core-975 specifications. http://ark.intel.com/, 2010.

[16] J.M. Rabaey et. al. *Digital Integrated Circuits*. Prentice Hall, second edition, 2003.

[17] R. Ho et. al. The future of wires. *Proc. of the IEEE*, 89(4):490–504, 2001.

[18] S. Perri et. al. A low-power sub-nanosecond standard-cells based adder. In *Proc. IEEE Int. Conf. Electronics, Circuits and Systems (ICECS)*, 2003.

[19] W. J. Dally et. al. Stream processors: Progammability and efficiency. *ACM Queue*, pages 52–62, 2004.

[20] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45nm design exploration. In *Proc. Int. Symp. Quality Electronic Design (ISQED)*, pages 585–590, 2006.

[21] T. Brunschwiler et. al. Forced convective interlayer cooling in vertically integrated packages. In *Proc. Intersociety Conf. Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, pages 1114–1125, 2008.

[22] The international technology roadmap for semiconductors (ITRS), 2009.

[23] A. Shubat. Manufacturing: Managing cost and risk. Semico Summit, 2009.