Modeling the Computational Efficiency of 2-D and 3-D Silicon Processors for Early-Chip Planning

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

*Centre for Microsystems Engineering, Faculty of Applied Sciences,
Lancaster University, Lancaster LA1 4YR, United Kingdom.
{m.grange,d.pamunuwa,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

Hierarchical models from physical to system-level are proposed for architectural exploration of high-performance silicon systems to quantify the performance and cost trade offs for 2-D and 3-D IC implementations. We show that 3-D systems can reduce interconnect delay and energy by up to an order of magnitude over 2-D, with an increase of 20-30% in performance-per-watt for every doubling of stack height. Contrary to previous analysis, the improved energy efficiency is achievable at a favorable cost. The models are packaged as a standalone tool and can provide fast estimation of coarse-grain performance and cost limitations for a variety of processing systems to be used at the early chip-planning phase of the design cycle.

1. INTRODUCTION

High-performance silicon processors will soon enter an era of massively parallel Tera-scale computing. Teraflops of computation inherently implies Terabytes per second of memory bandwidth. Higher-speed I/O and logic to drive the memory and computation towards the Tera-scale regime will result in prohibitively high power densities on a silicon area which is already limited by yield, cost and unfavorable interconnect RC delays [1]. Performance alone is already no longer acceptable as a metric, maximizing the performance per watt is the key to the successful continuation of Moore's law. 3-D integration holds promise to reduce interconnect power by eliminating long global wires and reducing off-chip I/O transactions whilst providing low-latency interconnections between stacked heterogeneous IP with electrically-fast through silicon vias (TSV). Furthermore, yield can be increased by partitioning a large silicon system over multiple layers and the integration of legacy dies can reduce the effort and cost involved with re-designing logic structures for new processes.

As design complexity increases with each generation of processors, more emphasis must be spent on the planning stages of the product design. For any new technology or architecture to become viable, it must conclusively demonstrate significant value at an acceptable risk and cost. To quantify the trade-offs between the many design choices available in the early chip-planning phase of a silicon-based processor, we have created hierarchical models to extract the performance limitations and the computational efficiency of 2-D and 3-D ICs under realistic cost and physical constraints. We base our models for computation on the underlying physical tenets of the three fundamental operations of any digital processing system: computation (logic), storage (memory), and communication (interconnect). Using these models and the accompanying tool the following features in single-chip processor design can be examined:

- **Scaling:** We model planar wires, devices, logic and memory from 180 nm down to 17 nm to extract the performance per Watt for 2-D and 3-D systems up to 16 layers as technologies scale. We also consider the effect of TSVs scaling from 20 μm down to under 1 μm. As an example we show that scaling the 65 nm Intel 80 core processor to 17 nm will increase the computational efficiency by 80%.
- **Architecture:** System-level design choices are captured by parameters in our model for computation including the layer partitioning, on-chip memory distribution, interconnect sharing ratio, memory technology (DRAM, Flash, SRAM), memory locality, data width and the ratio of on to off-chip transactions. We show that by partitioning the 65 nm Intel 80 architecture over four 3-D layers, an equal performance-per-Watt can be achieved as the same system developed in 45 nm.
- **System Constraints:** Specific systems can be modeled under the physical constraints of thermal, power, area, and frequency to aid in the optimization of a given topology for maximum performance under set of realistic application constraints. Using our models, we quantify the optimal number of layers for a processing system under thermal and power budgets dictated by common computing applications.
- **Yield and Cost:** We model and compare the variable and fixed costs for 2-D and 3-D processor architectures as a function of feature size and provide the necessary tools to assess the system costs. We conclude that for large silicon systems a 3-D implementation can recover total cost at lower volumes than a 2-D implementation despite the added expense of stacking.

The rest of this paper is organized as follows: we first discuss related work in section 2. In section 3 we introduce and explain the derivation of our performance model including its physical basis. We then in section 4 explore the design space by varying architectural parameters to demonstrate the flexibility of our model to determine performance limitations of both 2-D and 3-D processor topologies. In section 4.1 we examine a concrete processing system under physical constraints and discuss the cost of different implementations. Section 5 discusses the applications of our models and finally we end with our conclusions in section 6.

2. RELATED WORK

There are a number of available software suites that can provide

early chip-planning estimates, where many firms have in-house early-estimation tools. Cadence InCyte [2] for example allows a global view of a system including area, cost, performance and power using IP from manufacturers but is limited to 2-D designs and is restricted to specific applications. CACTI [3] provides an integrated environment for timing, power and area modeling, but is restricted to memory modeling. The authors of [4] offer predictive models of devices and 2-D wires, but no system-level analysis tool. 3-D ICE [5] and HotSpot [6] both offer thermal estimation of 3-D chip stacks but are not intended to model the underlying hardware features which generate the power profile. Our contribution is in providing a set of hierarchical models to analyze general 2-D and 3-D systems under physical, performance, and cost constraints before detailed knowledge of the design is known.

There has been an explosion of research over the past decade pertaining to 3-D integration both from a manufacturing and system-modeling perspective. The authors of [7] provide an excellent review of 3-D stacking technology and its benefits to high-end processors. In [8] the authors conclude that stacking DRAM on a single processor is a viable solution to overcome the increasing performance gap between memory and logic. The authors of [9] demonstrate that a 12-core stacked IC with no L2 cache outperforms an 8-core 2-D system with a large on-chip L2 cache by about 14% while consuming 55% less power. A similar performance result for stacked DRAM and logic was arrived at in [10]. However, none of these works present a general model to encapsulate the global physical architecture to enable design space exploration. In [11] the authors stipulated that the performance advantage of 3-D over 2-D ICs would shrink as feature size decreased below 65 nm, however recent work [12] analyzing the performance of a 16-core stacked IC and our own study has shown that in fact the advantage of 3-D ICs grows in lower technology nodes. An outlook of processor power and cooling strategies for 2-D and 3-D ICs as feature size scales is given in [13]. A number of thermal management strategies such as thermal via floorplanning, interposers and microchannel liquid cooling between die layers have been proposed to mitigate increased power density in 3-D ICs. In [14] cooling strategies are explored for multi-tier 3-D CPUs and test structures fabricated to compare parallel plate, microchannel, and pin fin structures in multiple configurations. Costing of a multi-tier 3-D processor is discussed in [15], where the authors show that different partioning strategies can result in lower cost for 3-D ICs. As a proof of concept, a 64-core stacked IC with 47,940 TSVs was fabricated to provide 63 GB/s of memory bandwidth at 277 MHz under a 47^{o}C thermal window [16].

3. MODELING PERFORMANCE

To introduce our contribution we look back to a metric first introduced by T. Claasen in [17] which describes the absolute maximum computational capability of a silicon system with a term called the *intrinsic computational efficiency* (ICE). The ICE of a system is calculated by filling the entire silicon area with the most fundamental computational circuitry, in this case 32-bit adders. Figure 1 shows Claasen's original ICE projection as feature size decreased and our own projection extending his work down to 17 nm. Two multi-core high-throughput processors are shown to exemplify the difference between a realistic system and the ideal upper bound of the ICE metric, which does not account for the I/O, interconnect, memory or general purpose control logic.

Our model attempts to build upon the ICE metric by factoring into account the expense of interconnect, memory, and I/O by introducing several design dependent parameters reflective of architectural features in a realistic processing system. To include the

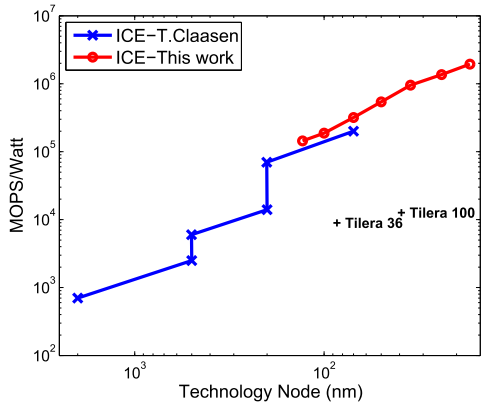

Figure 1: The Intrinsic Computational Efficiency of Silicon, where the the entire die area is filled with 32-bit adders. The difference between T. Claasen's projection and ours lies in the architecture of the 32-bit adder, leading to a minor difference in the energy/operation

effects of memory and interconnect, we first define the principle parameters of a computational system which enable us to model a variety of purpose-built processors.

- **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.

3.1 Effective Computational Efficiency

We define the *Effective Energy* (*EE*) for a 32-bit addition as:

$$EE^{\text{tn}}_{\text{arch}} = E^{\text{tn}}_{32} + \mu_T(E_{onchip} + E_{offchip}) \quad (1)$$

where the energy for on-chip and off-chip memory transactions is calculated from

$$E_{onchip} = \omega(e_1 + \Delta \times \text{E\_int}^{\text{tn}}_{\text{arch}}) \quad (2)$$

$$E_{offchip} = (1-\omega)(e_1 + \text{E\_int}^{\text{tn}}_{\text{arch}} + \text{E}_{\text{IO}})) \quad (3)$$

for a given technology node, tn, and a given architecture, $arch$, $\{2D, 3D2, 3D4, 3D8, 3D16\}$. 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 either over a 1 mm horizontal bus or from one vertical level to the next via a set of TSVs.
- E_{IO} is the energy to read/write the off-chip memory. It includes the energy consumed in the I/O drivers, the inter-chip communication as well as the memory controller.

The purpose 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 3D4 stack 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* in (1) we define the *Effective Computational Efficiency (ECE)* as $ECE^{tn}_{arch} = \frac{1}{EE^{tn}_{arch}}$ which gives the amount of computation achievable within the energy envelope of 1 Joule; or the amount of computations per second that can be carried out within the power envelope of 1 Watt. Figure 2 shows our model for computational efficiency as technology node scales. We have compared several recently implemented processors to demonstrate our model's correlation to real-world systems. Clearly, general-purpose processors exhibit lower *ECE* than domain-specific processors such as DSPs due to the large overhead of control to implement the required features in a desktop processor. The abstract system-level parameters in our model allow for any processor architecture to be represented.

3.2 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.

$$EA^{\text{tn}}_{\text{arch}} = A^{\text{tn}}_{32} + \mu_S\ \omega\ a_1\ + \sigma\ \text{A\_int}^{\text{tn}}_{\text{arch}} \quad (4)$$

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 [24] and between $8F^2$ and $4F^2$ for DRAM cells [25], where F is the minimum feature size.
- $\text{A\_int}^{\text{tn}}_{\text{arch}}$ is the interconnect area required for transporting a 32-bit word to memory (a 1 mm long 32-bit bus for 2-D systems or the area of 32 TSVs for a 3-D IC).
- σ is the interconnect sharing factor. If σ=1, no sharing takes place and every operator has its own, private interconnect across the system. If σ=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 [19] with 64 cores has 8 32-bit operators per core. For 512 operators 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.

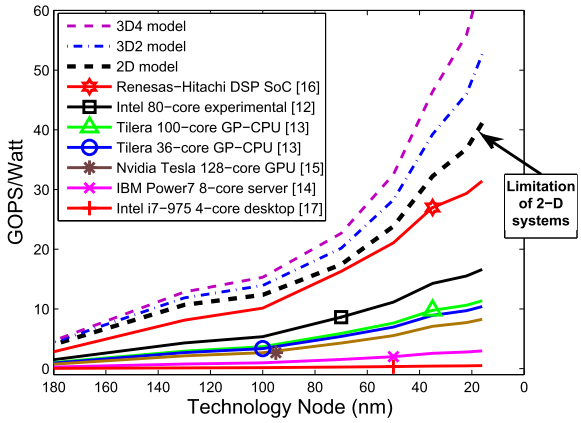

Figure 2: The Effective Computational Efficiency of recent multi-core processors [18, 19, 20, 21, 22, 23] (where markers indicate actual performance data) compared to our model. The limitation of 2-D systems shown by our model for 2-D computational efficiency is for a specific set of parameters (memory locality, bus width etc.). The model parameters can be altered to represent virtually any system, where the difference between our ideal and the actual implementation is dictated by the efficiency of the control circuitry.

3.3 Scaling circuits, devices and interconnects

To provide a firm foundation for our system-level model, we based our work on realistic circuit-level parameters extracted from published data, SPICE simulations, parasitic extraction tools, and consistent scaling methodologies [26]. The technology parameters cover global planar 2-D wires, TSVs, logical operations, memory transactions, thermal properties, transistors and leakage power.

3.3.1 Planar 2-D Wire and TSV Models

The minimum feature size on a die scales by roughly 0.7 each generation, however global on-chip wires do not scale as aggressively as intermediate or local wires [1]. We scale the global wires in CMOS ICs for technology nodes 180 nm down to 17 nm using a similar methodology to the authors of [27] 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 17 nm.

We have conducted field solver simulations of cylindrical, copper-filled TSVs to extract the relevant RLC parasitics. For the purpose of extracting parasitics and subsequent analysis, a representative structure for a TSV is assumed to be a copper-filled via with 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 [28]. The dimensions vary depending on the technology node; the cross-section is assumed to be uniformly circular, with 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. We use the extracted parasitics with the same methodology as the planar wires, where the bus width and switch factor match the 2-D parameters. Driver and receiver energy is also considered.

3.3.2 *Logical Operation and DRAM Scaling*

We model various logic operations such as a 32-bit addition or SRAM read, by using published data [29, 30] for a particular technology node and scaling the energy and area for future or past generations. Dally in [30] publishes the energy per add operation of a 32-bit adder in 130 nm 1.2 V technology as 5 pJ. A reasonable approximation, including leakage, for the energy and area in other technology nodes can be obtained according to well-practiced scaling methodologies [26] based on transistor feature size, supply voltage and thermal-leakage dependencies of transistors.

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 [25] 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 17 nm to DDR3. The flexibility in our memory model can allow for any type of memory to be integrated in the stack such as Flash or SRAM with a simple modification of the transaction energy.

3.3.3 *Power, Leakage and Thermal Models*

As the power density in 3-D ICs increases linearly with the number of active die layers, the thermal performance is an essential design consideration and subsequently forms a staple component of our physical 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. The FloTHERM® Computational Fluid Dynamic (CFD) solver was used to verify the accuracy of the models under a JEDEC still-air thermal test environment. We extracted leakage power trends across technology nodes using SPICE simulations with Predictive Technology Models [4] 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. Leakage is of especial concern with 3-D systems (particularly for lower technology nodes) due to the non-uniform temperature distribution across dies in the package, leading to a large variation of leakage power and hence reducing the overall maximum thermal ceiling and performance. 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.

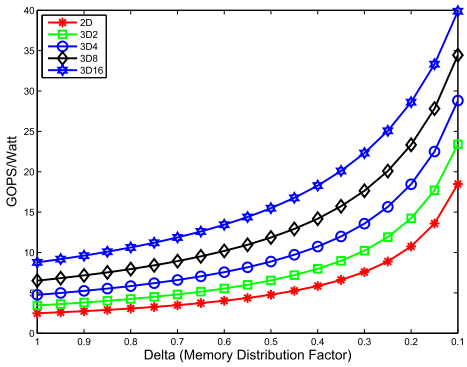

Figure 3: The effect of on-chip memory locality on the *ECE*, where decreasing Δ from 1 to 0 reduces the interconnect distance between operator and memory

4. DESIGN SPACE EXPLORATION

Varying the parameters (further described in section 3) 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 3 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 (see Figure 4). 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 17 nm technology. Hence, even if a 3-D topology can mitigate the cost of centralized memory, it is still growing exceedingly as 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 [19] 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.

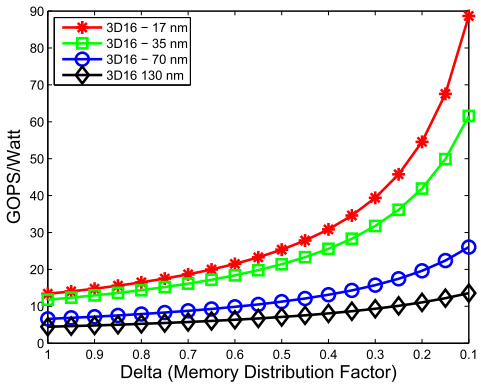

Figure 4: The effect of on-chip memory locality on the *ECE* for decreasing feature size in a 16-layer 3-D system

4.1 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 17 nm processor. However, when cooling limitations are considered, we find that 3-D ICs above eight layers will struggle to dissipate more than 15 W with conventional air-cooled heatsinks, where stacks of up to four layers may 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^{o}C 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 [14] 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 this particular 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 the 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 theoretical maximum performance 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 17 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. 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. To demonstrate this, Table 1 details the maximum operations per second (in GOPs) that a topology can achieve for a particular application constrained by temperature and power. 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, the optimal partitioning of the computational systems dictated by performance is clearly within the domain of 3-D integration. This is particularly important as low-power applications often require small footprints and integration of disparate technologies.

Table 1: The maximum operations per second (GOPs) for a 50 nm 400 mm^2 processor invoked in 2-D and 3-D topologies for different applications constrained by the maximum power under a thermal envelope. The maximum performance is shown in bold-face font.

App.	P(W)	T(^{o}C)	2D	3D2	3D4	3D8	3D16
Mobile	5	65	164	201	240	**357**	257
Laptop	25	75	1054	1295	**1471**	715	289
Desktop	65	85	**2445**	2383	1668	822	321
Server	150	120	**2867**	2797	1977	965	402

4.2 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

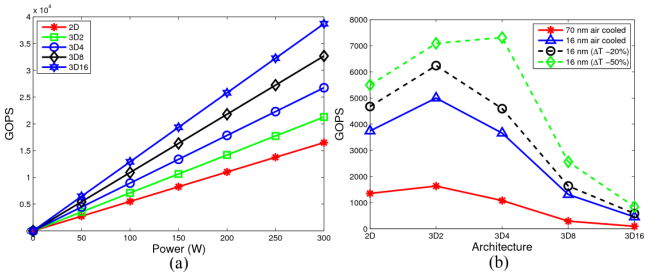

Figure 5: (a) The maximum throughput (GOPs) for our topologies constrained by power and (b) The maximum throughput with a thermal ceiling of 100^{o}C and 100 W. Dashed lines show additional heat removal beyond conventional air cooling

stack), generally acknowledged to drop off as the 4^{th} power of die area [26]. 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*.

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 [1] 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 [31, 1] 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, 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 [1], 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

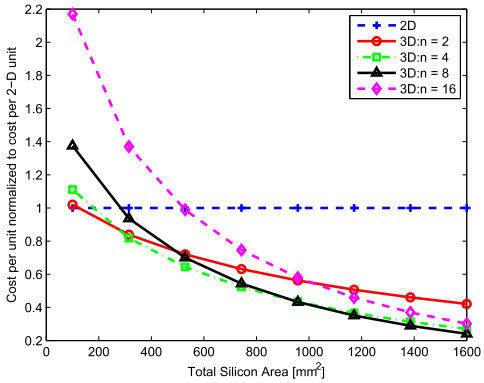

Figure 6: Die-to-Wafer variable cost

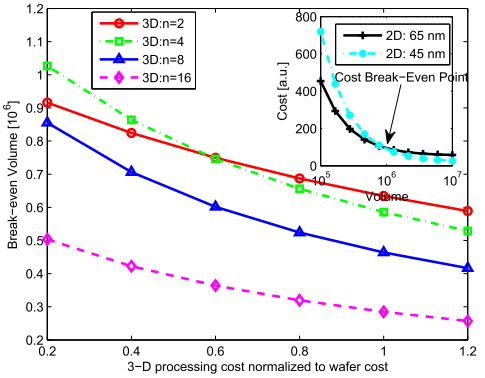

Figure 7: Variation of break-even volume with 3-D process cost for an NRE design cost increment of 75% in a technology shift

portion of the NRE cost as being 75% higher when moving from 65 nm to 45 nm [32], which results in Figure 7. 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 45nm 2-D implementation becomes more cost-effective than a 65nm 4-layer stack.

5. APPLICATIONS OF THE MODEL

Our hierarchical models for performance reflect global system-level design choices and are not intended to model specific processing systems, which would require detailed information of the control, logic, layout and application. The models we have created are intended to provide an early outlook of performance and cost for a wide-range of system design choices and technologies, which can aid in the partitioning and architectural organization of a processing system before the design begins. Similar works and in-house tools can provide fine-grained system-specific performance, area and cost information but are limited to that particular design space

Table 2: Equivalent model parameters for the Intel 80 Core

	Parameter	Intel 80 Core	Equivalent
N	# of Layers	1 (2-D)	1-16 (3-D)
A	Die Area	12.64×21.72 mm	275 mm^2/N
tn	Tech. node	65 nm	180-17 nm
b	Data width	32-bit	32
μ_s	Memory/Operator	2K SRAM/2 FPU	1KB/Op
μ_t	Memory/Operation	App. Specific	1-3
σ	Bus Sharing Ratio	8x10 mesh/160 FPU	18/160=0.11
Δ	Memory Distribution	NoC Mesh	0.01-0.1
ω	On/off-chip mem	All on-chip	1
P	Power (W)	20-230	App. Specific
T	Temperature(^{o} C)	80	80-100

and do not provide a general model for computational efficiency for 2-D and 3-D devices. Quantifying the trade offs between the investment in feature size reduction or 3-D stacking must be done for general systems and not necessarily for one particular design.

As an example, we model the Intel 80 core "teraflops" research chip to demonstrate the ability of our tool to model similar implemented processors. The 80 core device was designed to demonstrate the potential of a tiled multi-core processor design under a desktop power envelope. The 12.64 mm × 21.72 mm 65 nm die was packed with an 8×10 packet-switching network-on-chip mesh to interconnect the 80 tiled-resources consisting of two high-performance floating point units and 2KB of data SRAM. The processor achieved its peak performance per Watt (19.4) at 394 GFLOPs. Given this information we can populate a list of equivalent input parameters for our model of the processing unit as shown in Table 2. These parameters are then used to examine the performance limitations of that system under various constraints controlled by the user such as the number of layers the system can be partitioned into, the locality of the memory to the computational units, the technology node and TDP limitations. This provides a fast first-estimate of the capabilities of a particular system under different technological and architectural design choices which would normally be considered in the planning phase of a new product. This output, coupled with the cost comparison between 2-D and 3-D designs detailed in section 4.2, can allow fast and early optimization of a processing system for a given application, streamlining the planning stages and increasing emphasis for design and implementation of the device.

Shown in Figure 8 is our model for the Intel 80 core processor as compared to the same system partitioned over 2 (3D2) and 4 (3D4) layers. The reduced interconnect energy caused by partitioning the system in the vertical dimension demonstrates significant improvements in the overall energy efficiency. We show that older technology nodes can achieve a similar performance to an area reduced 2-D device if implemented in 3-D and that the performance increasingly favors 3-D topologies as feature size reduces. This plot provides a technological comparison both in feature size reduction and 3-D integration without adjusting any of the architectural parameters in the model. Further examination of the memory architecture or interconnect organization can provide essential insight into the performance limitations of a wide range of applications.

The underlying physical parameters on which our global system-level model is founded upon are well-defined and they alone can provide fast and accurate design space explorations. For instance, models for TSVs, 2-D planar wires, and logical operations are available and accurate for a wide range of geometrical input parameters. Our global methodology for assessing processor performance is an amalgamation of all of the underlying physical models in order

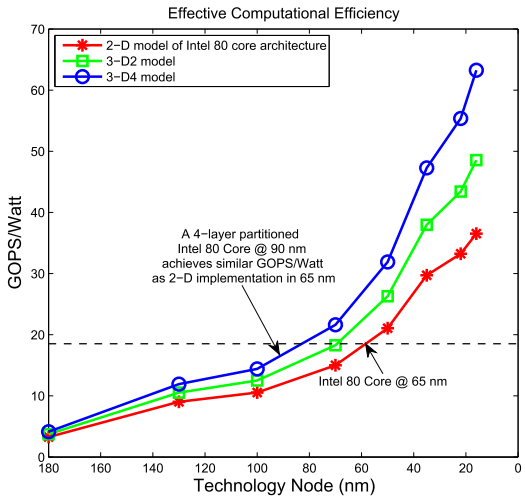

Figure 8: The computational efficiency in GOPs/Watt of the Intel 80 core floating point mesh-based processor as a function of feature size. We model the processor given the parameters in Table 2 in 2-D and 3-D implementations. The plot quantifies the efficiency of partitioning a similar system to the Intel processor in 3-D layers of 2 and 4 dies. The same performance per Watt can be achieved in lower technology nodes in 3-D and that the efficiency gap only increases as feature size reduces.

to provide a complete application perspective and quantify performance between different processor designs. We have packaged all of our models as an openly available web-enabled tool with range of user input parameters.

6. 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 17 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 [22] in Figure 2.

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 over-provided 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.
- A four-layer 3-D system can provide higher *ECE* than a 2-D system up to two technology nodes lower.
- 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.

The models which we have developed can be used to provide an early estimate of the performance limitations and capabilities of various processing systems before fine-grained layout and technological details are known. The quantification of both performance and cost for 2-D and 3-D systems as well as accurate parasitic models for a wide range of through silicon vias and 2-D wire geometries can provide designers with the framework to make realistic comparisons between the overwhelming number of CMOS design choices available in early-chip planning phase.

7. REFERENCES

[1] The international technology roadmap for semiconductors (ITRS), http://www.itrs.net, 2009.

[2] Cadence InCyte chip estimator, http://www.chipestimate.com, 2011.

[3] P. Shivakumar and N.P. Jouppi. Cacti 3.0: An integrated cache timing, power, and area model, 2001.

[4] 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.

[5] I. Beretta. A Mapping Flow for Dynamically Reconfigurable Multi-Core System-on-Chip. *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, 30(na):1–14, 2011.

[6] W. Huang et al. Hotspot: A compact thermal modeling method for CMOS VLSI systems. In *IEEE Transactions on*. Citeseer, 2006.

[7] G. H. Loh et. al. Processor design in 3D die-stacking technologies. *IEEE Micro*, 27(3):31–48, 2007.

[8] C.C. Liu et. al. Bridging the processor-memory performance gap with 3D ic technology. *IEEE Design and Test of Computers*, 22(6):556–564, 2005.

[9] T. Kgil et. al. Picoserver: using 3D stacking technology to enable a compact energy efficient chip multiprocessor. *SIGPLAN Notices*, 41:117–128, 2006.

[10] B. Black et al. Die stacking (3D) microarchitecture. In *Proc. IEEE/ACM Int. Symp. on Microarchitecture*, pages 469–479, 2006.

[11] P.D. Franzon et. al. Design and CAD for 3D integrated circuits. *In Proc. Int. IEEE/ACM Design Automation Conf.* (DAC), pages 668–673, 2008.

[12] K. Nomura et. al. Performance analysis of 3D-IC for multi-core processors in sub-65nm CMOS technologies. *In Proc. Int. Symp. Circuits and Systems (ISCAS)*, pages 2876 –2879, 2010.

[13] W. Huang et. al. Interaction of scaling trends in processor architecture and cooling. In *Proc.IEEE Int. Semiconductor Thermal Measurement and Management Symp. (SEMI-THERM)*, pages 198–204, 2010.

[14] 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.

[15] J. Zhao et. al. Cost-aware three-dimensional (3D) many-core multiprocessor design. In *Proc. IEEE/ACM Design Automation Conf.*, pages 126–131, 2010.

[16] M. B. Healy et al. Design and analysis of 3D-MAPS: A many-core 3D processor with stacked memory. In *Proc. Int. Custom Integrated Circuits Conf. (CICC)*, pages 1–4, 2010.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

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

[28] R. Weerasekera et. al. Compact modelling of through-silicon vias (TSVs) in three-dimensional (3-D) integrated circuits. In *Proc. IEEE Int. Conf. on 3D System Integration (3D IC)*, 2009.

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

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

[31] 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.

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