An RDL-Configurable 3D Memory Tier to Replace On-Chip SRAM

Marco Facchini\textsuperscript{12}, Pol Marchal\textsuperscript{1}, Francky Catthoor\textsuperscript{1}, Wim Dehaene\textsuperscript{2}

\textsuperscript{1}IMEC – Interuniversity MicroElectronics Center, Kapeldreef 75, B-3001 Heverlee, Belgium
\textsuperscript{2}ESAT-MICAS Katholieke University Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium

Abstract—In a conventional SoC designs, on-chip memories occupy more than the 50% of the total die area. 3D technology enables the distribution of logic and memories on separate stacked dies (tiers). This allows redesigning the memory tier as a configurable product to be used in multiple system designs. Previously proposed dynamic re-configurable solutions demonstrate strong dependence between read latency and dimensions of the mapped memory, leading to potential performance limitations. In this paper we propose a one-time configurable memory tier designed to minimize the performances overhead due to the commodity. Flexible configuration is enabled by smart memory macros and I/Os organization and a customizable redistribution layer routing. With respect to the dynamic re-configurability, the proposed design offers up to 40% faster access time, while saving more than 10% of energy per access. In addition production cost trade offs are analyzed.

I. INTRODUCTION

For all but maybe the leading companies (with dominant market share), SoC integration has turned into a dog market. Many of these companies have insufficient market share or work on products for which the cost benefits of scaling are unclear (such as HDTV, game processor chips). As a result, these companies struggle to launch a new must-have product every 18 months and face hard competition from those who still can. To retain scaling benefits, companies must increase manufacturing volumes, through merging with other suppliers or by developing a product portfolio based on common components/platforms.

Emerging 3D-IC technologies support the latter option. By extending the SoC philosophy, they enable the integration of a system using standard building blocks manufactured in separate dies. In this paper, we pursue the above idea, proposing to separate the design in a custom logic tier and a configurable memory tier (see Figure 1). Most of the on-chip SRAMs in the conventional SoC design are migrated to this configurable memory tier. The memory tier consists of many small memory banks (grains), which can be configured for and reused across multiple systems. If volumes for this memory tier can be made sufficiently high, replacing embedded SRAM with configurable memory tier results in a lower cost than monolithic SoC integration (see section V). The overhead of configurability and the additional 3D integration cost can be further amortized with technology node scaling and/or by customizing the process technology of the memory tier e.g. 4-metal layers rather than >6 typically necessary for logic.

In the proposed concept the memory grains are configured using the redistribution layer (RDL), necessary for 3D integration (Figure 2).

TABLE I. BENEFITS OF THE RDL-CONFIGURABLE DESIGN OVER THE RESPECTIVE 2D-DESIGN AND OTHER COMMODITY MEMORY DESIGNS \cite{1}

<table>
<thead>
<tr>
<th>2D-design</th>
<th>Commodity design of \cite{1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>delay</td>
<td>may not be achievable</td>
</tr>
<tr>
<td></td>
<td>same potential as us</td>
</tr>
<tr>
<td>power</td>
<td>up to -40%</td>
</tr>
<tr>
<td>cost</td>
<td>up to -9%</td>
</tr>
<tr>
<td></td>
<td>-8%</td>
</tr>
</tbody>
</table>

In the next section we introduce the 3D technologies that are being considered in this paper, and describe advantages and limitations of previously published works on commodity memories. The proposed RDL-configurable commodity memory design is introduced in Section III. Section IV presents the analyzed scenarios and the performance/cost estimation methodology. Finally, in Section V, we explain the experimental results obtained before conclusions and future work. (Section VI).

II. RELATED WORK

A. Commodity enabling technologies

Through-silicon vias and \(\mu\)-bumps enable 3D integration by creating a direct electrical link (3D-link) between dies arranged in tier. Their excellent electrical characteristics (Table II) permit inter-die
communication as fast as on-chip routes [2], allowing distribution of the SoC IPs over different dies (e.g.[3]).

A redistribution layer (RDL) may be required to align differently patterned tiers’ contacting I/Os (TSVs and/or passivation openings). An RDL is a thick plated metal layer with superior electrical characteristics vs. typical metal layers [4] (Table II). Today, RDLs are processed by the packaging houses rather than foundries. As a result, they are processed at relatively low cost, but with limited precision (pitch limited to 10 μm).

<table>
<thead>
<tr>
<th>TABLE II. CHARACTERSISTICS OF INTERCONNECT TECHNOLOGY [1][4][5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacitance (fF)</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>TSV [4][5]</td>
</tr>
<tr>
<td>μ-bump [1]</td>
</tr>
<tr>
<td>RDL [4]</td>
</tr>
<tr>
<td>Metal 2 - 7</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

‡ = estimation based on similar technology measurements
** = target/assumption
*** = approximated values from TSMC90nmG

B. SoC commodity memory design

The idea of a configurable 3D memory tier to replace the fast on-chip SRAMs was presented in [1]. In that work the memory tier consists of uniform memory cells connected between them with a switched-based network (Figure 3). 3D gate-ways interconnect the switch network with the logic tier. Each IP can be interconnected with a number of memory cells from a single gateway by configuring the switch network. After reconfiguration, each IP may access only that given set of memories (or domain), propagating its own clock frequency.

![Figure 3. Re-configurable memory die proposed in [1]. Uniform memory grains (b) interconnected with a switch-based network (c and d). This network causes extra and variable access delay penalty (a) (TSMC90nmG).](image)

The performance of the commodity memory is limited by the switch network. In practice, a signal may need to traverse several switch boxes before reaching the targeted memory cells. The accumulated delay limits the maximum access frequency of the configured memory.

E.g. assuming the switch-box described in Figure 3c and 3d, the access delay to different memory cells varies between 0.62ns and 3.04ns depending on the number of switch-boxes traversed. This limits the short term applicability of configurable memories, as the system design needs to be adapted to the extra delay (e.g., additional pipelining).

III. ONE-TIME CONFIGURABLE COMMODITY MEMORIES (THROUGH RDL)

In comparison, the approach presented in this paper produces minimal and uniform extra access delay by interconnecting the memory grains on the tier using the RDL, rather than the switch-boxes.

Using the RDL to create appropriate short-circuits between the memory grains I/Os, we can configure them to work as a single unit with a wider word-width or a larger memory space.

In the next section we describe the proposed memory tier floor-plan, necessary to mitigate the performance and flexibility limitations imposed by the 10μm pitch of the RDL.

A. Commodity floor-plan matched with today’s RDL

The geometrical data presumed in table II constrain the possible distributions of the memories I/Os on the die layout. E.g. arranging the TSVs (or passivation openings) in a matrix where the distance between the elements is equal to the minimal pitch (as in [1]), does not allow to route RDL traces between the outer TSVs, thus isolating the I/Os enclosed inside the matrix. Extra spacing and/or multiple RDLs hardly solve the problem.

We propose a floor-plan where all the I/Os of each single grain are arranged in a minimum-pitch mono-dimensional array. This array is in general wider than the memory grain itself, thus a number of grains are placed side by side, as well as the relative I/Os (Figure 4b). The grains interface does not need to be identical as long as their I/Os can be aligned to enable orthogonal RDL routing (see next section).

![Figure 4. Minimum-pitch array placement (b) enables free RDL routing across adjacent memory I/Os (a). The 4x8-grains commodity die (a) is configured as four separate memories. Each separate route has to be connected to a μ-bump (not-shown).](image)

The I/Os arrays are connected to the respective memory grains utilizing traditional technology metal routing layers. In the example of figure 4b, and for the considered TSMC90nmG technology, the four routing metal layers required for memory manufacturing are sufficient to avoid routing congestions. Also the entire routing can be confined in the I/O arrays area. Buffering is required to reach the I/Os at the ends of the arrays, with consequent performance loss (see section V). Multiplexing/serialization solutions [1] may help to reduce I/Os count but it would further compromise on power consumption.

B. Customizing commodity memory routing on a single RDL

In Figure 4a we show a memory die populated with 4x8 memory grains with homogeneous characteristics. In addition we show the RDL routing configuring the die in 4 heterogeneous memory units.

All the memory grains in a single memory unit receive the same address, command and clock signals. Consequently the respective I/Os of the different grains are short-circuited together. This is not always the case for the data I/Os, which can be independently interconnected...
to the logic die (parallel grains connection) or also short-circuited (series grain connection). The data I/O connections determine if the memory unit has a larger than the memory grain data bit-width (e.g. red mapped area in Figure 4a) or addressing space (e.g. blue mapped area). In the latter case only one of the memory grains has to be active at the time. For this purpose it is possible to encode the different bank-enable signals with the output of a bank-address decoder placed on the logic die. μ-bumps (not shown in Figure 4a) have to be created on each exposed RDL trace for later bonding with the logic die.

As this is a commodity product, one may not need all the resources available. Therefore, some I/Os could be not connected and isolated during RDL deposition, while also some memory grains may be left unused.

RDL routing is always possible in the direction orthogonal to the I/O arrays as far as we want to short-circuit memories. Indeed, it is not possible to route over the I/Os with a single layer RDL without creating a short-circuit. RDL routing is also possible in the direction parallel to the I/O arrays. However, memory grain size limits the routing resources in this direction. Nevertheless one may still route the few address, command and clock signals (e.g. orange mapped area in Figure 4a) and eventually use the logic die routing resources for the data I/Os connections.

IV. CASE STUDY

A. Tier-arrangement scenarios

During the case study exploration, three different memory tier organizations are compared with reference 2D scenarios:

1) Custom memory tier. The memory macros are just lifted from the logic die to a dedicated memory die, generating a customized memory tier;

2) Switch-boxes (re-configurable) [1]. Logic switch boxes are integrated in the tier to provide run-time re-configurability;

3) Back-side metallization (RDL). This is one realization of the proposed RDL-configurable memory tier. In this case the RDL is created under the bulk of the silicon die to interconnect the TSVs.

For the three scenarios we consider the TSVs integrated in the memory tier, which eventually enables multiple memory tier stacking.

![Figure 5. Scenarios and corresponding interconnections electrical model](image)

B. Design case v.s. commodity tier characteristics

A representative subsystem of the AVC H263 has been used as a test-vehicle for evaluating the overhead of the different memory tier scenarios in terms of performance, power and total chip cost.

TABLE III. MAPPING THE MEMORY MACROS OF THE CASE STUDY DESIGN ON THE COMMODITY MEMORY. ALL MACROS HAVE A SINGLE ACCESS PORT

<table>
<thead>
<tr>
<th>Memory banks in the subsystem</th>
<th>Commodity mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td># banks</td>
<td># branches</td>
</tr>
<tr>
<td>4</td>
<td>2048</td>
</tr>
<tr>
<td>16</td>
<td>128</td>
</tr>
<tr>
<td>1</td>
<td>128</td>
</tr>
<tr>
<td>1</td>
<td>128</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
</tr>
<tr>
<td>1</td>
<td>128</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
</tr>
</tbody>
</table>

The subsystem has a total of 33 memory banks with different characteristics (Table III). Those are mapped on a commodity memory die populated with two kinds of memory grains having different addressing spaces while the same data word-width. The mapping has been performed according to Table III.

The data word-width is chosen to be 32bits as simulations with larger bit-width (64bits) show comparable system performances overheads. Also, the area overhead in the switch-box scenario grows exponentially with the word-width due to logic and routing multiplication. Contrary, narrower bit-width leads to considerable additional power overhead (>20% for 16bits).

The addressing space of the commodity memory grains has been chosen to obtain a good mapping of the given study-case design while minimizing the amount of different sized grains. This choice allows for a fair comparison between the different evaluated scenarios.

C. Estimation 3D performance

The different scenarios overhead is calculated using a datasheet based model (Figure 6). The model extracts delay, power and area estimations from the TSMC libraries based on the different sequences of standard cells and wire loads.

![Figure 6. Generating interconnects overhead](image)

The overheads are then included into the memory macro libraries as additional latency, hold/setup constraints, leakage and energy per access. Also, the libraries I/O pins information is modified to be the one of a properly buffered I/O.

This last modification gives us tool-free control on the scenarios specific buffering and overheads, decoupling the buffering analysis from the floor-plan generation. Thus, the design synthesized netlist can be processed just twice for placement and routing: one time for the 2D reference scenario and another one for all the 3D integration scenarios.

Timing violations and power estimations are performed by exchanging the libraries with the scenario specific ones (Figure 7).

V. COMMODITY MEMORY ADVANTAGES

In this section we show the benefits of the proposed RDL scenario compared to literature [1] and the reference 2D scenario. We also compare to results achieved by producing the RDL-configurable scenario in more advanced technology nodes.

This technology scaling is enabled by the high volume at which we can produce a configurable die, thus amortizing the large not-
recurring engineering costs of scaling. Finally, a system designer will benefit of the better performances coming with scaling, independently from its targeted final production volume.

### A. Technology scaling makes commodity cheap

For estimating the logic die cost (Table IV) we have been using the production cost of a typical fab processing for large volumes, applied to the cost model proposed in [7]. Also we included the TSV processing cost as a fixed price per wafer and the mask-set cost projected on a production of 20M dies. The cost assumed for the memory die is the same as for a commodity DRAM memory (.028$/mm² –market spot price for 3 metal-layer DRAM memory).

TABLE IV. PRODUCTION COST PER CHIP. INCLUDING THE EFFECT OF MEMORY DIE SCALING ON THE PROPOSED RDL SCENARIO

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Logic die</th>
<th>Memory die</th>
<th>Total cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D (reference)</td>
<td>0.358</td>
<td>---</td>
<td>0.358</td>
</tr>
<tr>
<td>Custom memory</td>
<td>0.293</td>
<td>0.108 ***</td>
<td>0.402</td>
</tr>
<tr>
<td>Switch boxes</td>
<td>0.293</td>
<td>0.111</td>
<td>0.405</td>
</tr>
<tr>
<td>RDL</td>
<td>0.293</td>
<td>0.104</td>
<td>0.397</td>
</tr>
<tr>
<td>RDL 65nm</td>
<td>0.293</td>
<td>0.079</td>
<td>0.372</td>
</tr>
<tr>
<td>RDL 45nm</td>
<td>0.293</td>
<td>0.045</td>
<td>0.338</td>
</tr>
<tr>
<td>RDL DG 45nm **</td>
<td>0.293</td>
<td>0.032</td>
<td>0.325</td>
</tr>
</tbody>
</table>

All values are expressed in $ per die

* = includes mask cost (.031$/ and 3D-bonding (.041$)

** = we use memory grain size with the double of the capacity

*** = price/mm² higher than commodity (same as logic but less metal layers)

The additional price for the 3D-bonding and the silicon area occupied by TSVs makes the 2D design cheaper than the custom 3D scenario (this would not happen for larger designs [8]). The RDL scenario results in less area occupation, thus cost than the switch boxes scenario.

Finally, the RDL scenario becomes economically more efficient than 2D design when the memory die technology scales to 45nm. Also the choice of the memory grain impacts the die cost as it is shown for a double data word-width case.

### B. Performance evaluation

Without considering scaling, the access to a memory bank on the commodity die is always slower than accessing the same bank on the logic die. This is due to the delay required to transit the 3D interconnect load (μBump, RDL and TSVs).

Table V shows the increasing performance advantage of the RDL-configurable scenario when using the switch based scenario when extending the memory unit addressing space. Indeed, increasing the addressing space results in a larger RC load in the first scenario, whereas the signals must be propagated through a longer series of switchboxes in the latter scenario, which has a larger latency penalty.

Applying the different scenarios to the AVC design case we didn’t observe any timing violations. This is because the memory transmission latency margins of this specific design case were larger than the additional scenarios overhead.

Also, the power consumption of the RDL-configurable design is shown to be better (>10%) then the switch-based design. In Figure 8 we show the total power consumption of the memory die for the different scenarios.

Figure 8. Total power consumption in the commodity memory tier for the considered design study-case. For this estimation we suppose a 0.2 toggling activity probability on the subsystem input pins.

With respect to equivalent on-chip memory traffic (Figure 8 ref), the RDL-configurable scenario (Figure 8.3) consumes 41% more power. This is partly due to the 3D interconnect load (14%) and to the commodity mapping of large data word-width memory requirements on several narrow word-width grains (~8%). The remaining of the power (~19%) goes into the routing and buffering required to connect the memory grains with the I/Os.

Power consumption may be further reduced with technology scaling. An SRAM in 65nm is expected to consume less than the 80% of the correspondent SRAM in 90nm. Thus using 45nm technology for the proposed RDL scenario (3) we end up consuming less power than the reference 2D scenario (Figure 8). As discussed above the use of advanced technologies for the commodity SRAM does not necessarily imply the use of these technologies also for the logic die.

### VI. Conclusions

In this paper we have described a low latency, power efficient and flexible 3D-stacked commodity memory tier. The design organization is suited to be customized using an inexpensive RDL layer. Simulations show up to 40% faster access time to the memory tier and 10% less power consumption in comparison to other solutions present in literature. Also the proposed design reveals to be economically advantageous in respect to the 2D case if scaling to 45nm is available for the commodity tier.

However the question remains open which is the optimal choice of the grain to maximize the applicability of the commodity memory in a maximum number of applications. This will be object of further investigation.

### REFERENCES


