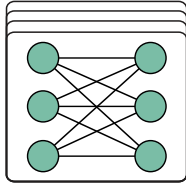




This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, Netherlands, Switzerland



Addressing the call/topic: H2020 ECSEL-2018-2-RIA
Research and Innovation Action



TEMPO

Technologies and hardware for neuromorphic computing

Deliverable

D4.3 – 3D DNN Module PnR

Work Package:	WP4 (Design and architecture)
Dissemination level:	Confidential
Official due date:	30 April 2020
Document editor:	Geert Van der Plas (IMEC)
Contributing partners:	IMEC
Internal reviewers:	Ilja Ocket (IMEC), Björn Debaillie (IMEC)
Document version:	V1

© Copyright TEMPO Project. All rights reserved.

This document and its contents are the property of the TEMPO Partners. All rights relevant to this document are determined by the applicable laws. This document is furnished on the following conditions: no right or license in respect to this document or its content is given or waived in supplying this document to you. This document or its content is not be used or treated in any manner inconsistent with the rights or interests of TEMPO Partners or to its detriment and are not be disclosed to others without prior written consent from TEMPO Partners. Each TEMPO Partner may use this document according to the TEMPO Consortium Agreement.

1. Publishable summary

Memory-on-Logic 3D system partitioning can help improve system performance in a great way. Benefits come from more direct interconnect between the logic and the memory macros, but also from the fact that standard cells in the logic die can be placed closer to each other, thus significantly reducing the contribution of the interconnect to the critical path delay.

Benefits from the 3D system integration are architecture dependent. However, 3D system integration is particularly well suited for Deep Learning, Artificial Intelligence, Neural Networks, and other architectures that use complex memory hierarchies that fully share data among different cores [1]. Such architectures put significant pressure on the interconnect requirements to the memory and very often must scarify capacity to allow reasonable performance [2].

By analysing 3D place and route of a state-of-the-art many-core architecture with 64 low-power RISC-V like cores we were able to show that the 3D performance could be improved by 40% compared to 2D (imec iN3). Power and area of the 3D system still need to be co-optimized together with the performance. This is because current 3D timing constraints, automatically generated for a given 3D partition use fixed timing budget for all 3D paths in the system. This causes excessive buffer insertion that skews power and area of the 3D SoC.

To reach such significant 3D performance benefits we need aggressive 3D integration technologies. In this work we have used hybrid Face-to-Face bonding at wafer level to allow sub 1 μm 3D structure pitch. In our design, and to allow signal and power routing we need a 3D pitch of 0.56 μm . This is driven by the fact that the partitioning of lower cache levels results in a significant 3D pin count over a small area. Partitioning of upper levels would certainly relax the pitch requirements, but will not bring as many performance benefits, since upper hierarchical levels are typically more tolerant to interconnect delay.