A Classification of Memory-Centric Computing

HOANG ANH DU NGUYEN, JINTAO YU, MUATH ABU LEBDEH, MOTTAQIALLAH TAOUIL, and SAID HAMDIOUI, Delft University of Technology
FRANCKY CATTHOOR, Inter-university Micro-Electronics Center (IMEC)

13

Technological and architectural improvements have been constantly required to sustain the demand of faster and cheaper computers. However, CMOS down-scaling is suffering from three technology walls: leakage wall, reliability wall, and cost wall. On top of that, a performance increase due to architectural improvements is also gradually saturating due to three well-known architecture walls: memory wall, power wall, and instruction-level parallelism (ILP) wall. Hence, a lot of research is focusing on proposing and developing new technologies and architectures. In this article, we present a comprehensive classification of memory-centric computing architectures; it is based on three metrics: computation location, level of parallelism, and used memory technology. The classification not only provides an overview of existing architectures with their pros and cons but also unifies the terminology that uniquely identifies these architectures and highlights the potential future architectures that can be further explored. Hence, it sets up a direction for future research in the field.

CCS Concepts: • **Computer systems organization** → *Special purpose systems*; • **Hardware** → *Spintronics and magnetic technologies*;

Additional Key Words and Phrases: Computation-in-memory, resistive computing, memory-centric computer architectures

ACM Reference format:
Hoang Anh Du Nguyen, Jintao Yu, Muath Abu Lebdeh, Mottaqiallah Taouil, Said Hamdioui, and Francky Catthoor. 2020. A Classification of Memory-Centric Computing. *J. Emerg. Technol. Comput. Syst.* 16, 2, Article 13 (January 2020), 26 pages.
https://doi.org/10.1145/3365837

1 INTRODUCTION

For several decades, technology scaling has provided a 43% performance gain for each successive node and cheaper computers as a result of a higher operating frequency and lower cost per transistor, respectively [15, 54]. On top of that, smart architectural improvements such as pipelining and cache hierarchies have increased computer performance up to 50% every 2 years [49]. However, CMOS scaling suffers from three main walls: leakage wall, reliability wall, and cost wall [45], while computer architectures also face three walls: memory wall, power wall, and instruction-level

The results presented in this article have been obtained in the framework of the project "Computation-in-memory architecture based on resistive devices" (MNEMOSENE), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780215.
Authors' addresses: H. A. D. Nguyen, J. Yu, M. A. Lebdeh, M. Taouil, S. Hamdioui, and F. Catthoor, Mekelweg 4, 2628 CD, Delft, the Netherlands; emails: {H.A.DuNguyen, J.Yu-1, M.F.M.AbuLebdeh, M.Taouil, S.Hamdioui}@tudelft.nl, Francky.Catthoor@imec.be.

1550-4832/2020/01-ART13 $15.00
https://doi.org/10.1145/3365837