Technische Universiteit
Eindhoven
University of Technology

A Parallel-Access Method for 3D-Stacked DRAMs

Manil Dev Gomony, Benny Åkesson and Kees Goossens
Department of Electrical Engineering / Electronic Systems

Problem statement

→ Devise different DRAM architecture solutions using 3D-Integration technology for improved bandwidth and evaluate the real-time performance in terms of memory efficiency and bandwidth.

Introduction

→ In 3D-Integration, multiple dies are stacked on top of each other and interconnected using Through Silicon Vias (TSV) [1].

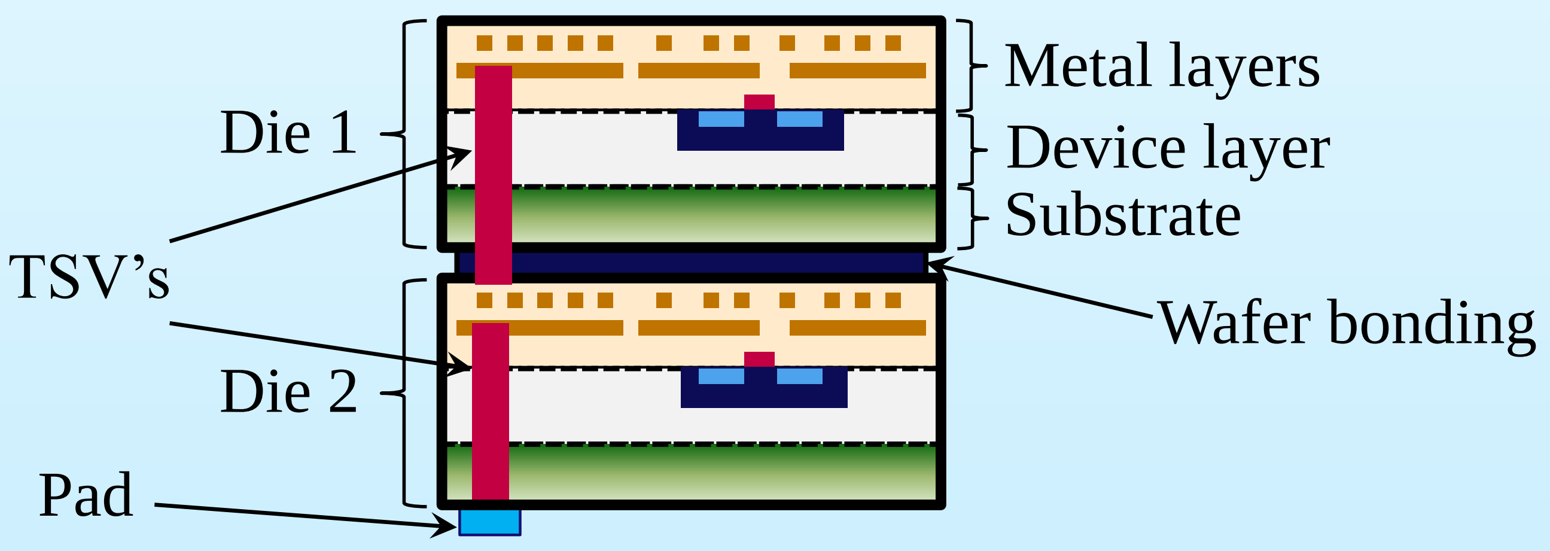


Figure 1: Two dies interconnected using TSVs

→ Low area requirements and excellent electrical characteristics of TSVs enable a large number of connections between the logic layer and DRAM [2].

Proposed architecture

→ We consider a *Parallel-Access (PA)* method in which each bank of the DRAM has its own command and data interface to the memory controller.

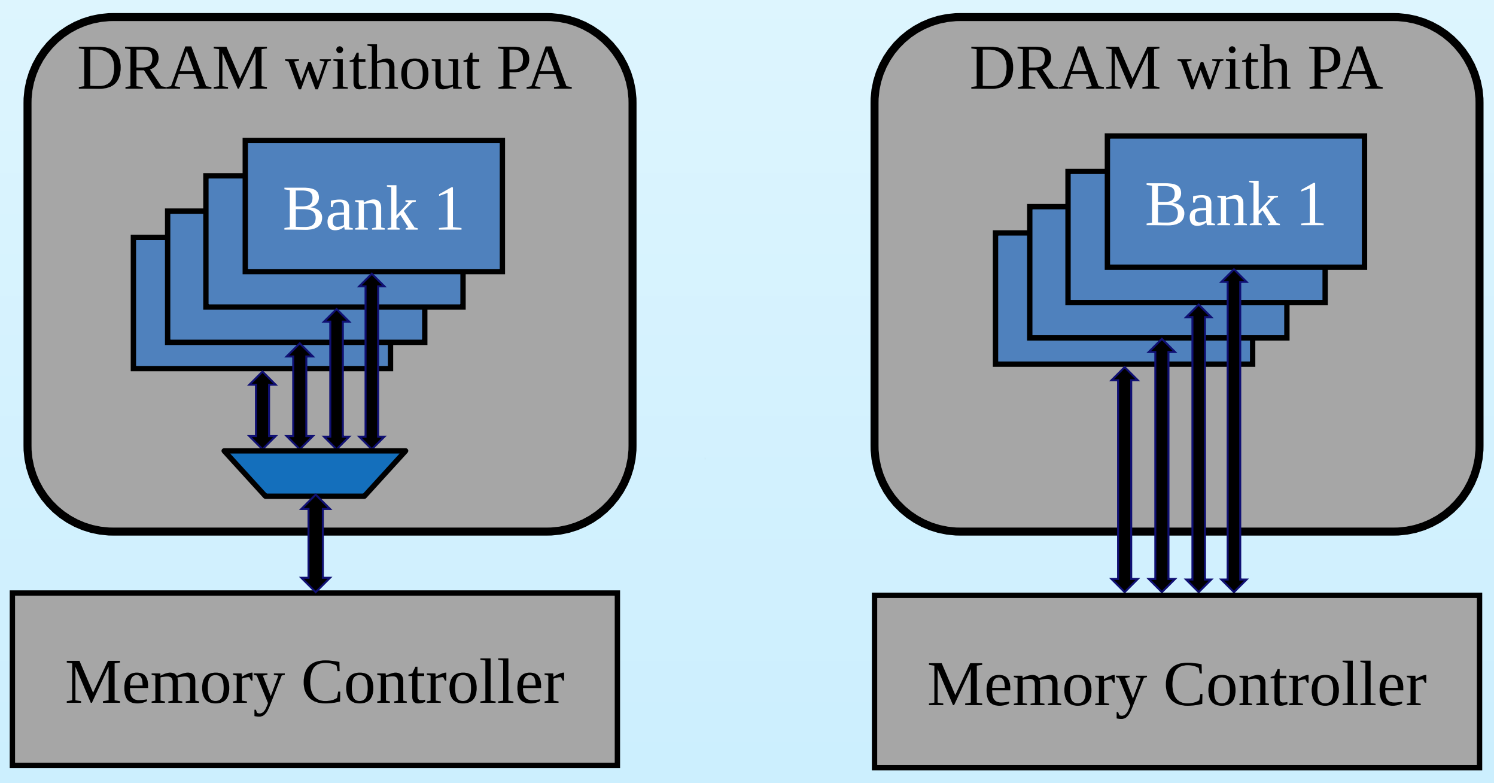
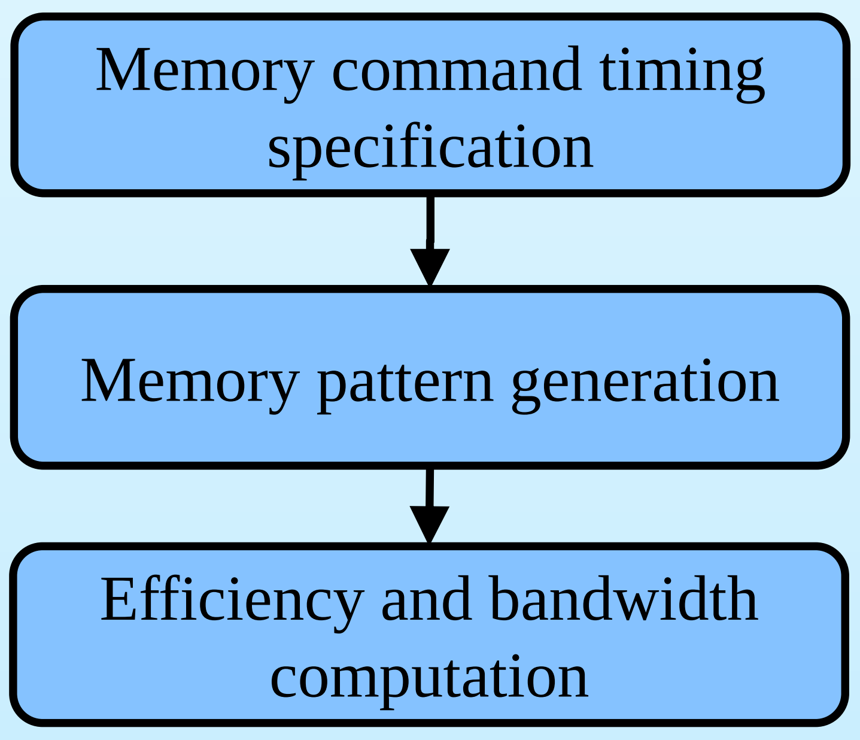


Figure 2: DRAM architecture without and with Parallel-Access

→ This approach is about using the 3D-Integration technology enabler to make banks into independent memories by removing shared logic meant to reduce the number of pins. This increases the provided bandwidth.

Methodology

→ Memory efficiency and real-time bandwidth is computed using *memory patterns*, which are pre-computed sequences of DRAM commands and that satisfies the DRAM command timing constraints [3].



Memory command timing specification

Memory pattern generation

Efficiency and bandwidth computation

Figure 3: Performance analysis method

→ memory efficiency = command efficiency × read-write efficiency × bank efficiency × data efficiency

→ peak bandwidth = operating frequency × io width × data rate

→ gross bandwidth = memory efficiency × peak bandwidth

Performance results

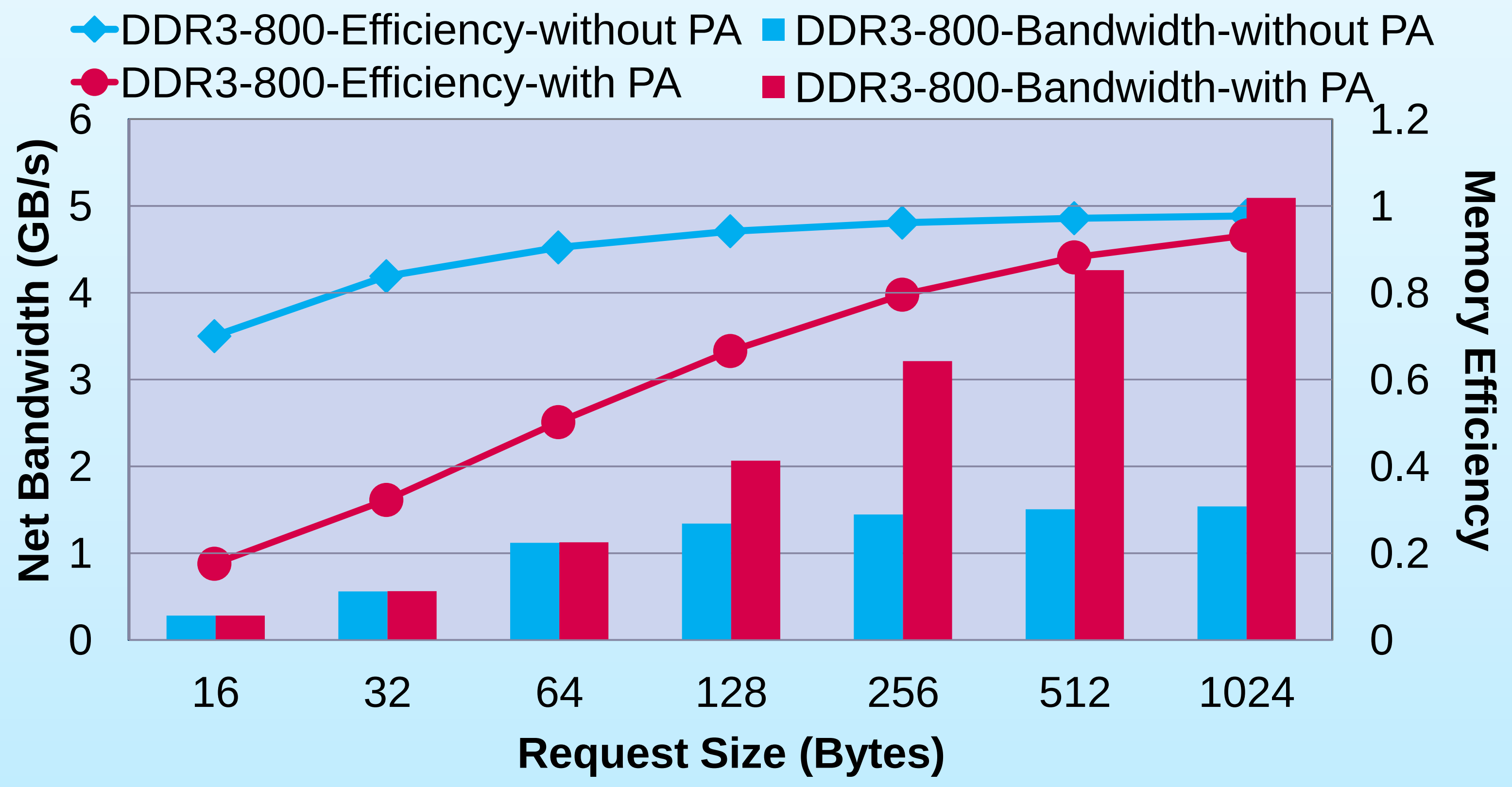


Figure 4: Net bandwidth and memory efficiency of DDR3-800 with PA and without PA for different request sizes

Conclusions

→ With smaller request sizes, parallel-access provides the same net bandwidth at the cost of lower efficiency because of its larger access granularity.

→ Parallel-access provides up to 3× gain in net bandwidth with larger request sizes.

→ Parallel access is suitable for applications with large request sizes and huge bandwidth requirements.

References

[1] Wei-Chung Lo et al., "3D Chip-to-Chip Stacking with Through Silicon Interconnects ." in Proc. VLSI-TSA, 2007.
[2] M. Facchini et al., "System-level Power/performance Evaluation of 3D stacked DRAMs for Mobile Applications." DATE, 2009.
[3] B. Akesson et al., "Classification and Analysis of Predictable Memory Patterns," in Proc. RTCSA, 2010.

Manil Dev Gomony
m.d.gomony@tue.nl

Acknowledgements
This work was supported in part by SenterNovem (an agency of the Dutch Ministry of Economical Affairs), as part of the EUREKA/CATRENE/COBRA project under contract CA104.