Fahad Bin MUSLIM, Kashif INAYAT, Muhammad Zain SIDDIQI, et al., 2025. SAPER-AI accelerator: a systolic array-based power-efficient reconfigurable AI accelerator. *Frontiers of Information Technology & Electronic Engineering,* 26(9):1624-1636. https://doi.org/10.1631/FITEE.2400867

SAPER-AI accelerator: a systolic array-based power-efficient reconfigurable AI accelerator

Key words: Artificial intelligence (AI) accelerators; Application-specific integrated circuit (ASIC) design; Systolic arrays; Low-power designs

Corresponding author: Fahad Bin MUSLIM
E-mail: fahad.muslim@giki.edu.pk
ORCID: https://orcid.org/0000-0002-4153-360X

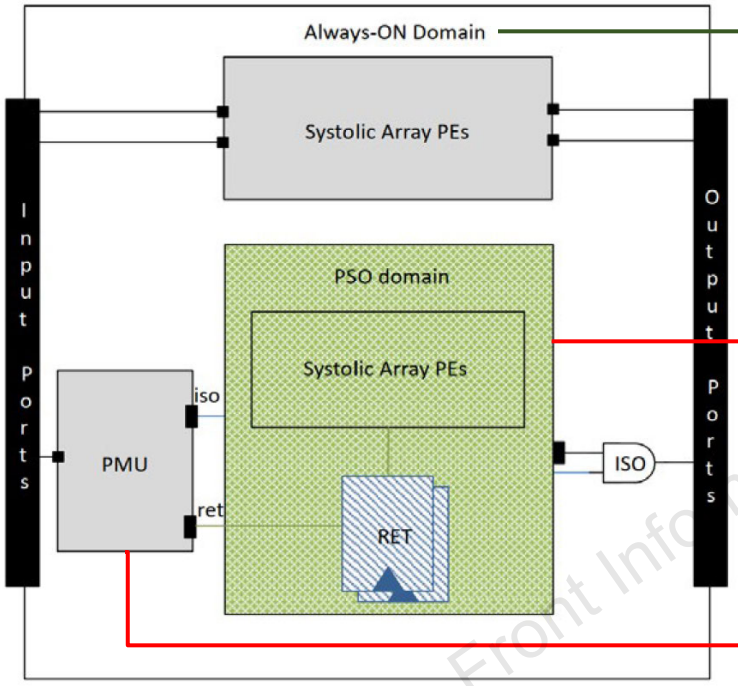
Motivation

- Systolic arrays (SAs) are known to efficiently accelerate dense deep neural network (DNN) computations, but modern networks with sparse and heterogeneous layers can cause many processing elements (PEs) to remain idle, thus leading to reduced energy efficiency.
- To overcome this, we propose coarse-grained power gating of idle PEs in a reconfigurable manner, thereby significantly improving energy efficiency without incurring expensive micro-architectural efforts.

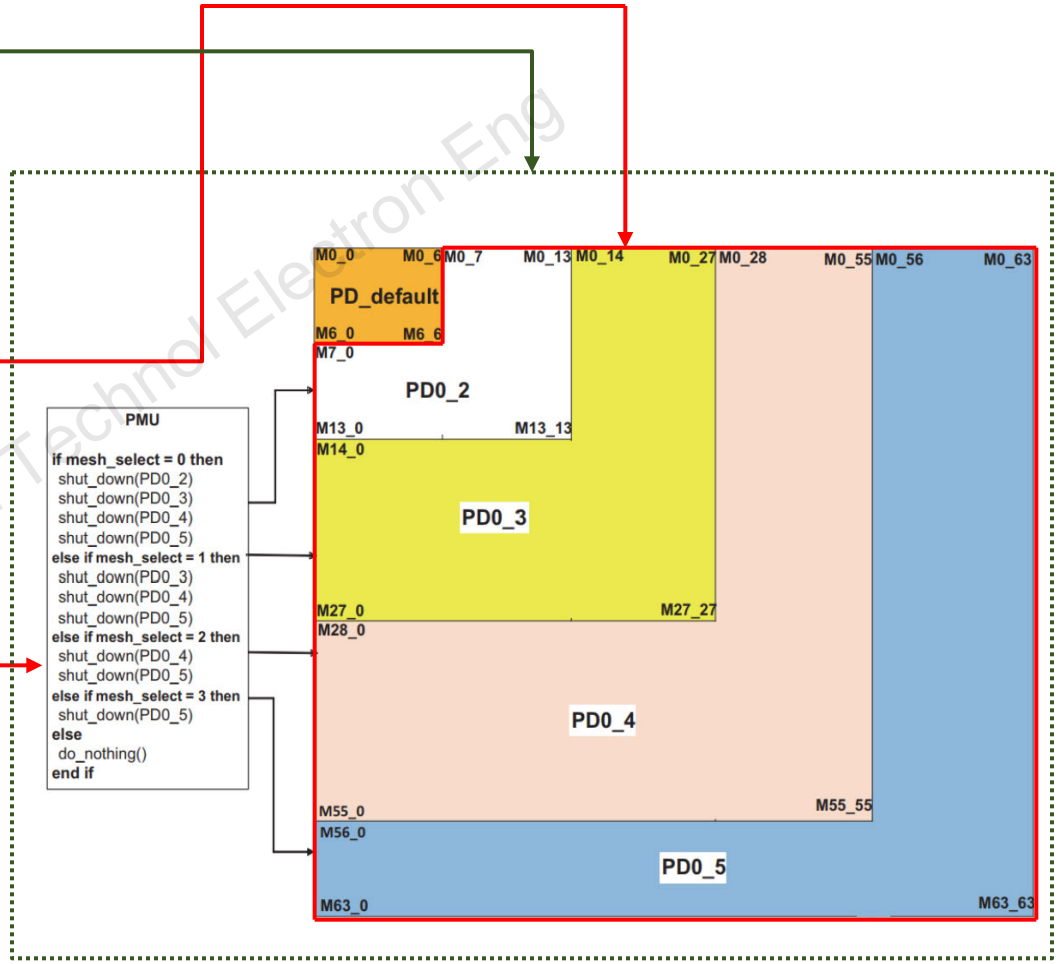
Main idea

- Reconfigurable 32×32 and 64×64 SA-based artificial intelligence (AI) accelerators are implemented to cater to the varying workloads of some DNN models, thereby ensuring the full utilization of the active PEs of the underlying SA designs at any given time.
- Coarse-grained power gating of row and column PEs is accomplished to match varying computational requirements of various layers of the DNN workload with the main aim to save power without expensive micro-architectural modifications.
- Realistic power analysis is performed by using actual DNN workloads, e.g., MobileNet and ResNet50 workloads.
- A detailed comparison is made based on power, performance, and area (PPA) and power delay product (PDP) parameters.

Methodology



Generalized power-aware HW design



Power-aware 64x64 SA abstract illustration

Design test cases

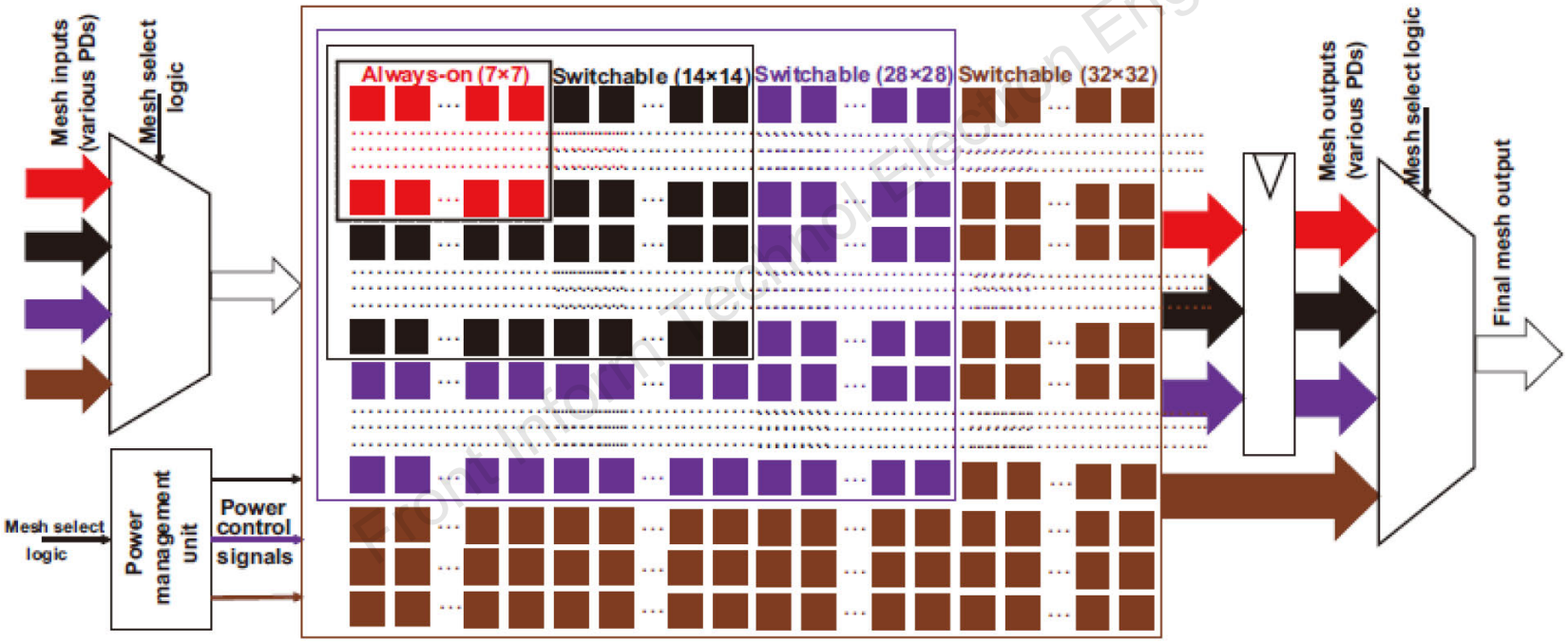


Fig. 5 Detailed view of the 32×32 SA with power control signals. References to color refer to the online version of this figure

Design test cases

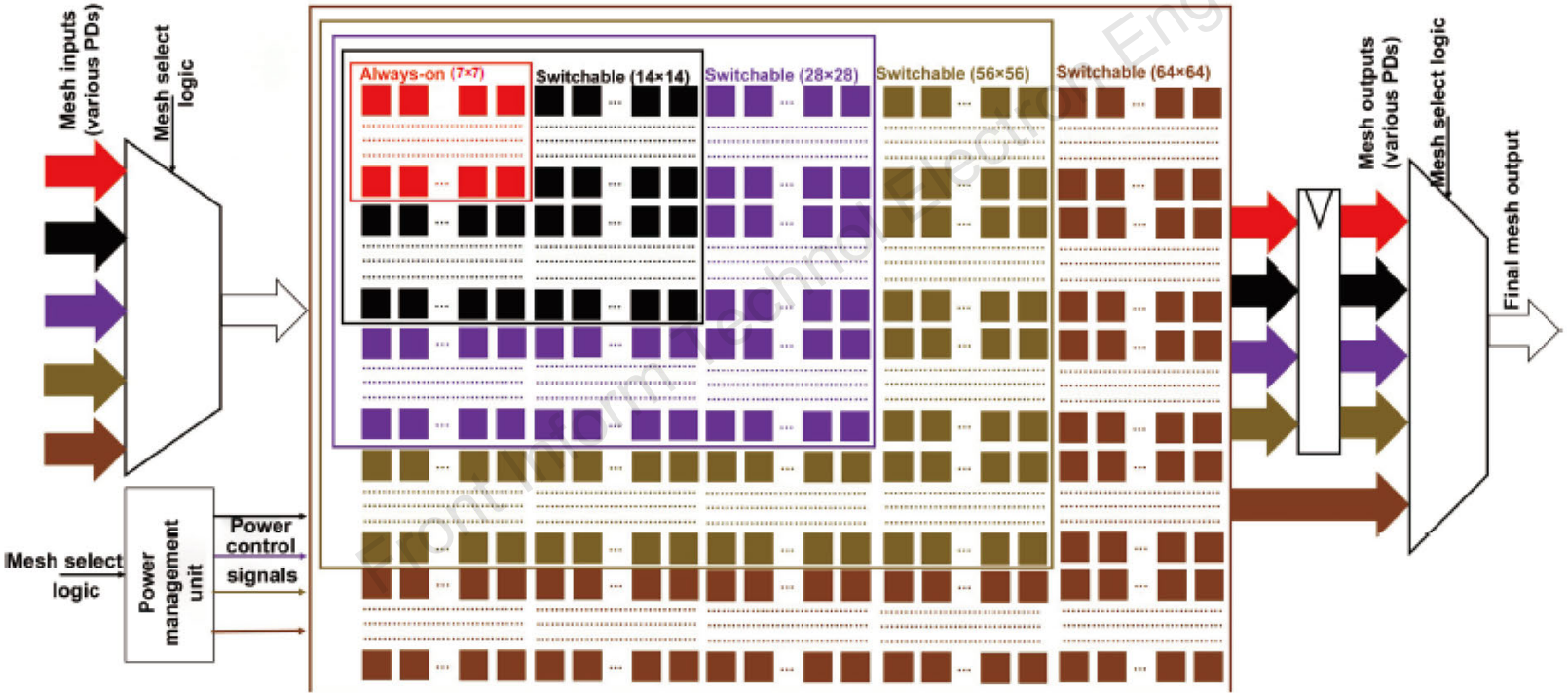


Fig. 6 Detailed view of the 64×64 SA with power control signals. References to color refer to the online version of this figure

Results

Table 5 Performance analysis of the 32×32 SA across DNN workloads with (w/) and without (w/o) UPF

Workload	Delay (ns)		Area (mm^2)		Power (W)		PDP (W·ns)	
	w/o UPF	w/ UPF	w/o UPF	w/ UPF	w/o UPF	w/ UPF	w/o UPF	w/ UPF
MobileNet	**1.56**	1.91	**7.74**	8.41	1.094	**0.980**	**1.71**	1.87
ResNet50	**1.56**	1.91	**7.74**	8.41	1.077	**0.950**	**1.68**	1.81

Better results are in bold

Table 6 Performance analysis of the 64×64 SA across DNN workloads with (w/) and without (w/o) UPF

Workload	Delay (ns)		Area (mm^2)		Power (W)		PDP (W·ns)	
	w/o UPF	w/ UPF	w/o UPF	w/ UPF	w/o UPF	w/ UPF	w/o UPF	w/ UPF
MobileNet	**1.56**	1.96	**30.8**	34.2	3.72	**2.90**	5.80	**5.68**
ResNet50	**1.56**	1.96	**30.8**	34.2	3.60	**2.70**	5.62	**5.29**

Better results are in bold

Results

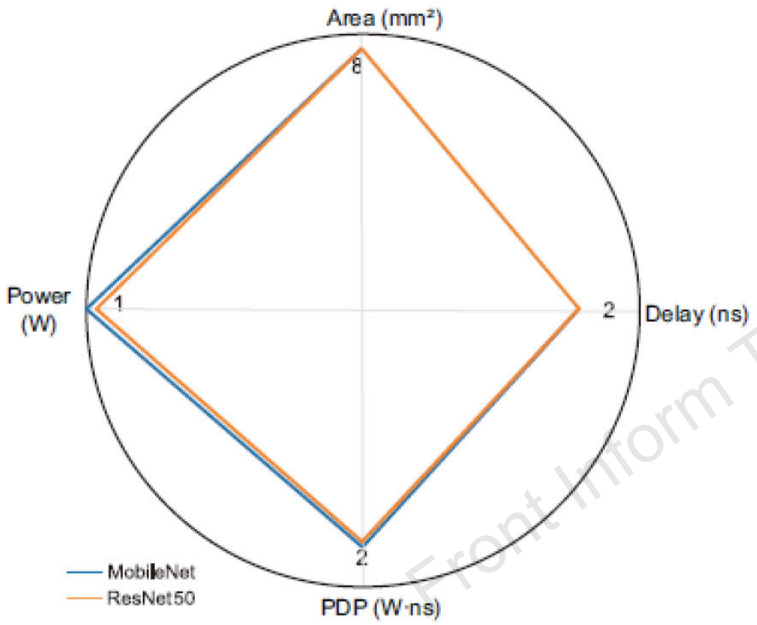


Fig. 7 Best-case performance analysis of the 32×32 SA design

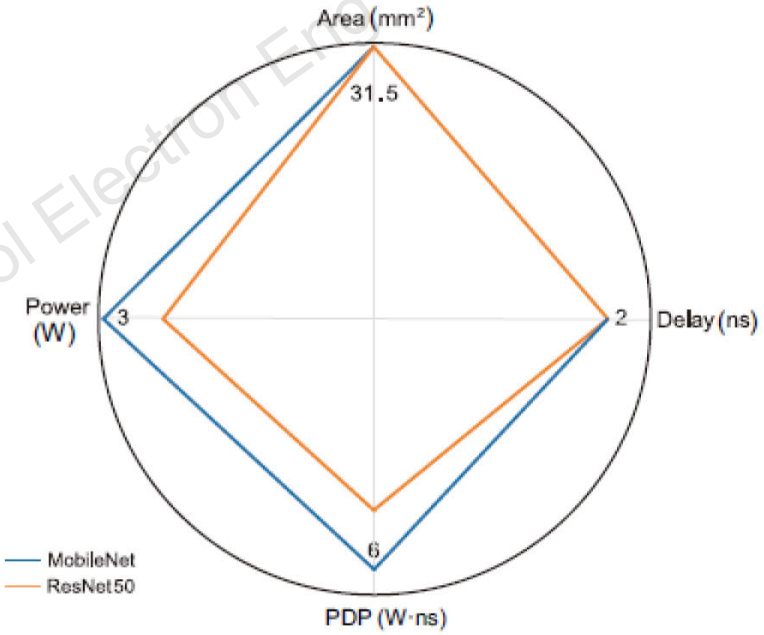


Fig. 8 Best-case performance analysis of the 64×64 SA design

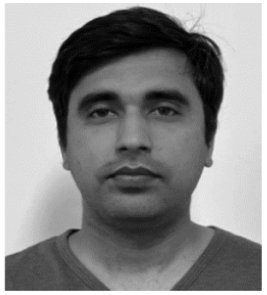
Conclusions

- Power optimizations cause area and delay penalty.
- Power efficiency is improved as the SA size increases.
- ResNet50 workload performance is better compared with MobileNet performance for larger SAs, because greater uniformity in the ResNet50 convolutions is more favored by the underlying SA architecture.

Fahad Bin MUSLIM is presently serving as an Assistant Professor of the Faculty of Computer Science and Engineering at GIK Institute, Pakistan. He received the M.S. degree in communication engineering and Ph.D. degree in electronics engineering from Chalmers University of Technology, Sweden and Politecnico di Torino, Italy, in 2010 and 2017, respectively. Previously, he was a researcher at the Incheon National University, Republic of Korea. His research interests include high-level synthesis, ML accelerators, and heterogeneous compute systems with the focus on power efficient architectures.

Kashif INAYAT received his B.E. degree in electronics engineering from Iqra University, Pakistan in 2014, his M.S. degree in electronics and computer engineering in 2019, from Hongik University, Republic of Korea, and his Ph.D. in electronics engineering from Incheon National University (INU), Republic of Korea in 2023. He worked as a researcher at the R&D Institute, INU, as a registered researcher with Samsung Research Funding and Incubation Center for Future Technology (SamsungFTF), Republic of Korea, and as a Senior Research Engineer at Barcelona Supercomputing Center, Spain. He is currently working as a Senior Engineer at Qualcomm Technologies Inc. Cork, Ireland. His current research interests include neuromorphic computing, ML accelerators, computer arithmetics, low-power digital design in ASICs/FPGAs, and hardware security.

Muhammad Zain SIDDIQI received his M.S. degree in electrical engineering from the COMSATS University, Islamabad, Pakistan, and his Ph.D. degree in electronics engineering from Tsinghua University, China. He is currently serving as an Assistant Professor at the Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute (GIKI), Pakistan. Prior to joining GIKI, he worked as a Lecturer at Lahore Leads University, Pakistan. His current research interests include reconfigurable intelligent surface, massive MIMO, mmWave communications, resource, power allocation in wireless networks, and green communication.



Safiullah KHAN received the Ph.D. degree in computer engineering from Gachon University, Republic of Korea, in 2023. He was a Project Engineer with the National Radio and Telecommunication Corporation, Haripur, Pakistan, for two years. He also remained a Research Fellow with the Centre for Secure Information Technologies (CSIT), Institute of Electronics, Communications and Information Technology (ECIT), Queen's University, Belfast, U.K. Currently, he is a Lecturer with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, U.K. His research interests include efficient hardware implementations of cryptographic protocols, post-quantum cryptography, and blockchain.

Tayyeb MAHMOOD is a CTO at Nextwave, Republic of Korea. He received the Ph.D. degree in information and communication engineering from the Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 2013. He also worked as a postdoctoral researcher at System on Chips (SoC) Lab, Incheon National University, Republic of Korea. Previously, he was working as an Assistant Professor with the Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan. His research interests include low-power computing, Internet of Things, and embedded systems.

Ihtesham ul ISLAM received the B.S. degree in computer system engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2006, the M.S. degree in electronics and communication engineering from Myongji University, Republic of Korea, in 2009, and the Ph.D. degree in computer and control engineering from the University of Politecnico di Torino, Italy, in 2015. From 2009 to 2012, he worked as a Lecturer with the FAST-National University of Computer and Emerging Sciences, Peshawar. From 2015 to 2020, he worked as an Assistant Professor with the Department of Computer Science and IT, Sarhad University, Peshawar. He is currently working as an Associate Professor with the National University of Sciences and Technology, Pakistan. His research interests include computer vision and machine learning.