V10: Hardware-Assisted NPU Multi-tenancy for Improved Resource Utilization and Fairness

Yuqi Xue Yiqi Liu Lifeng Nai† Jian Huang

† 

An Increasing Demand for Machine Learning Services in the Cloud

Microsoft Azure ML

Amazon Web Services

Google Cloud AI Platform

...

An Increasing Demand for Machine Learning Services in the Cloud

Conversational AI

Language Translation

Recommendation System

Computer Vision

...

Microsoft Azure ML

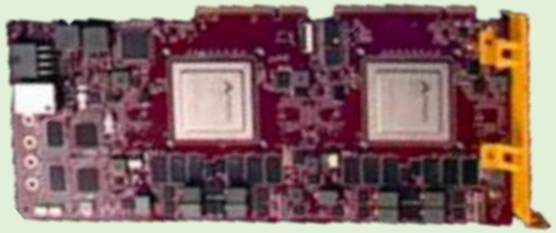
Amazon Web Services

Google Cloud AI Platform

...

Neural Processing Units Are Being Widely Deployed in Cloud Platforms

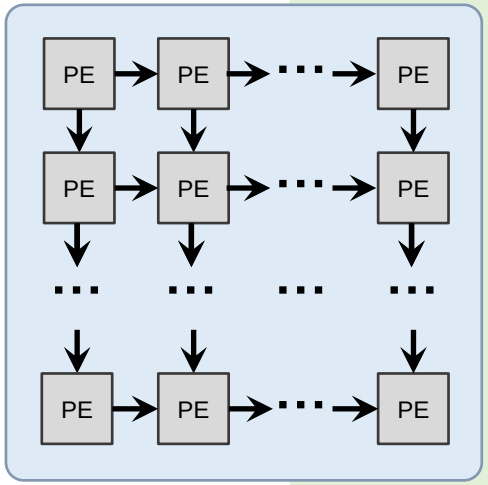
Google TPU



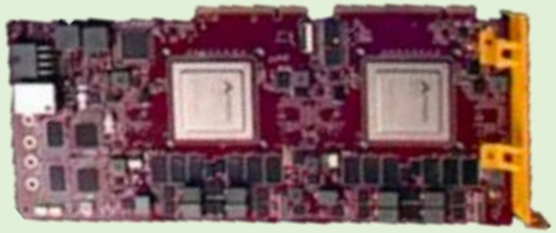
AWS Inferentia

...

Neural Processing Units Are Being Widely Deployed in Cloud Platforms



Google TPU

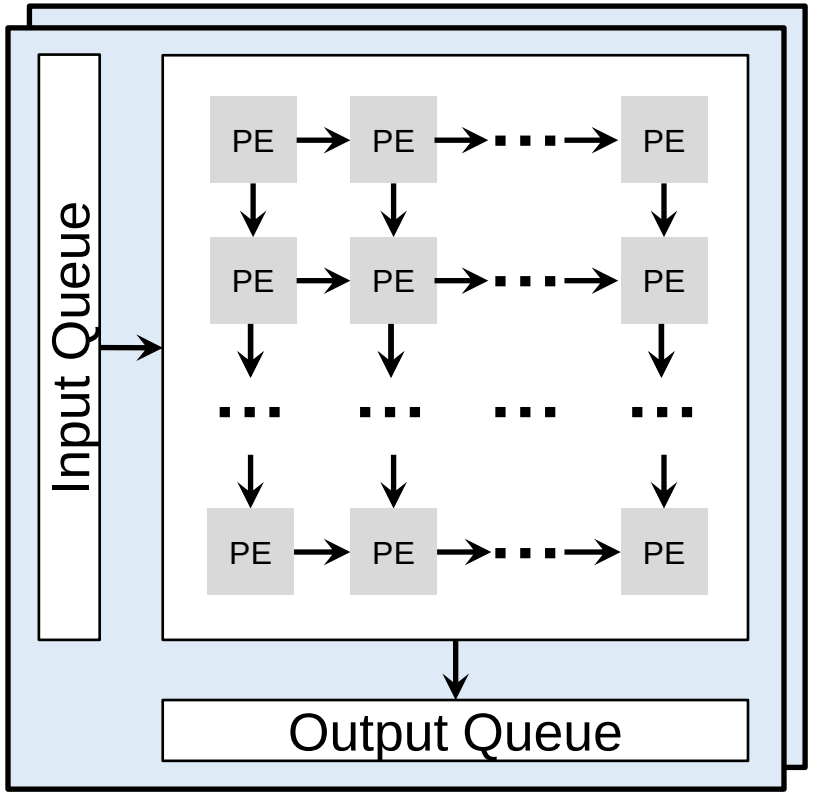


AWS Inferentia

...

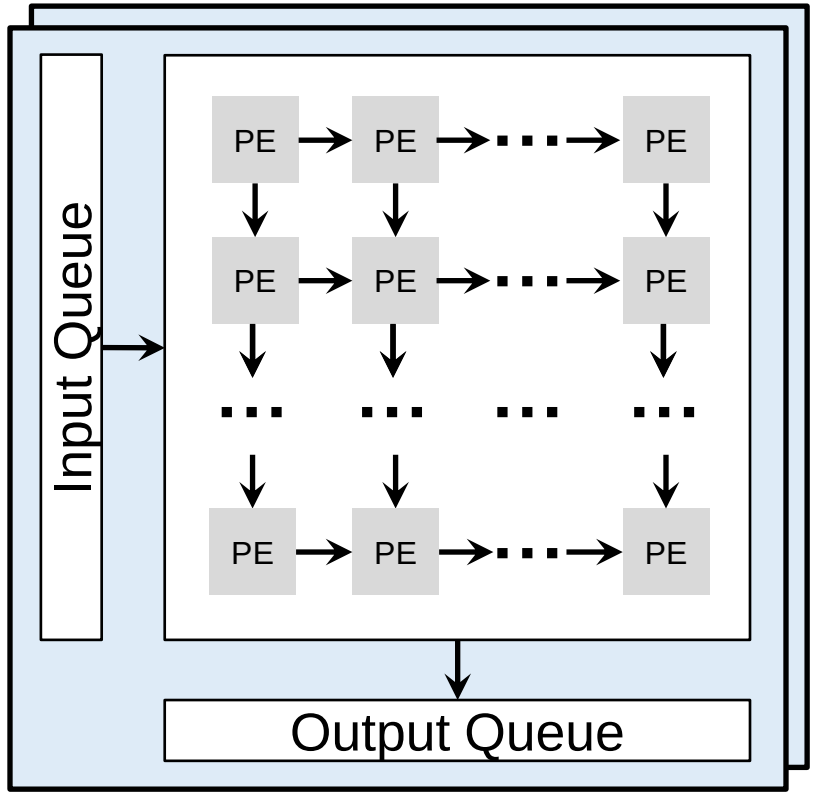
Many Neural Processing Units Employ Systolic Arrays

Neural Processing Unit 101



Systolic Arrays (SAs)

Neural Processing Unit 101

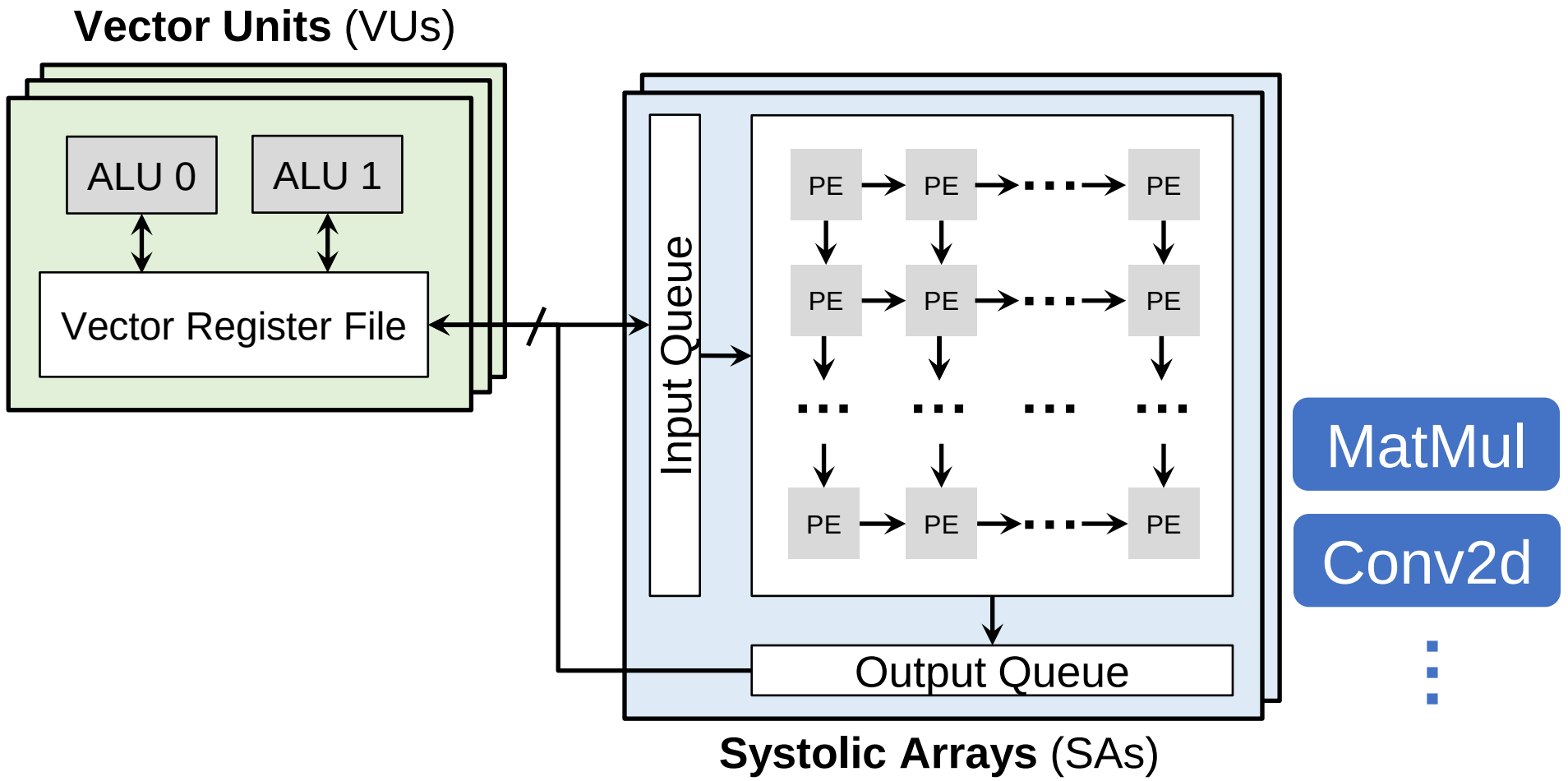


Systolic Arrays (SAs)

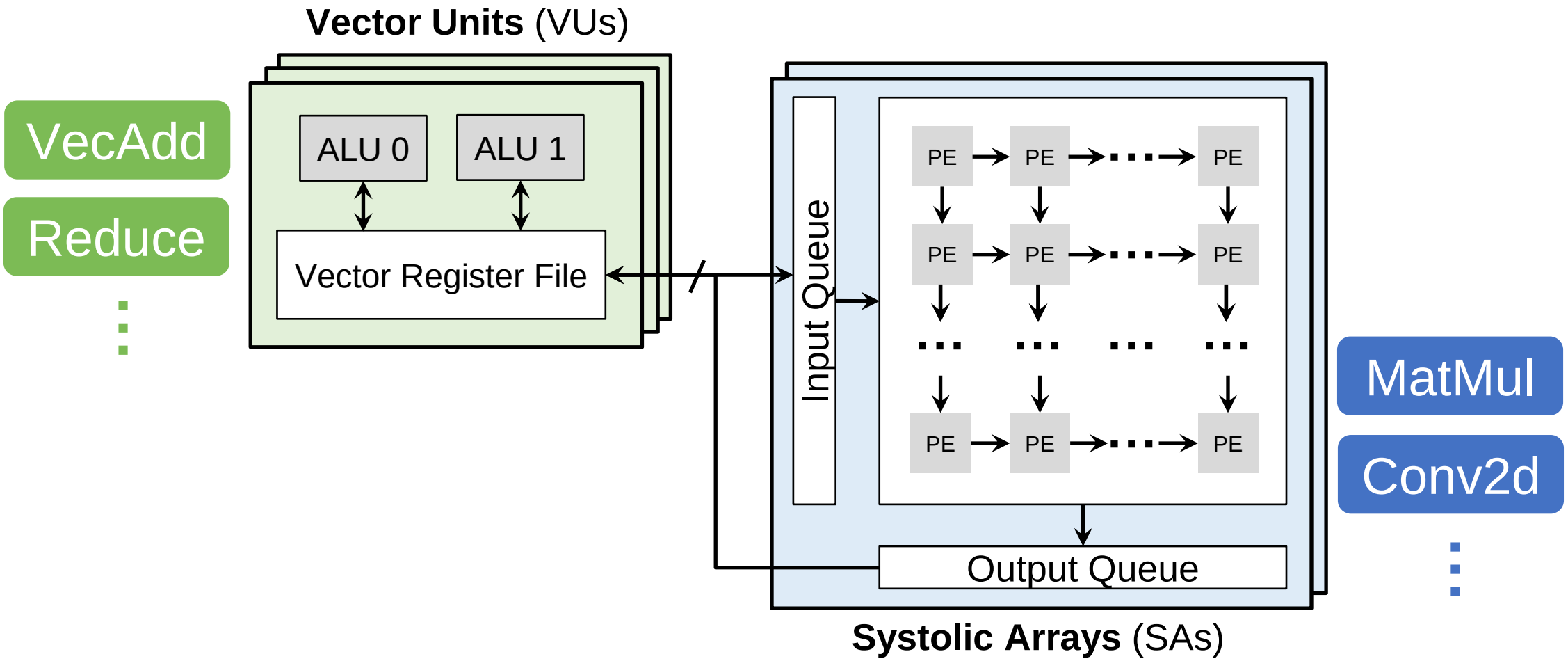
MatMul

Conv2d

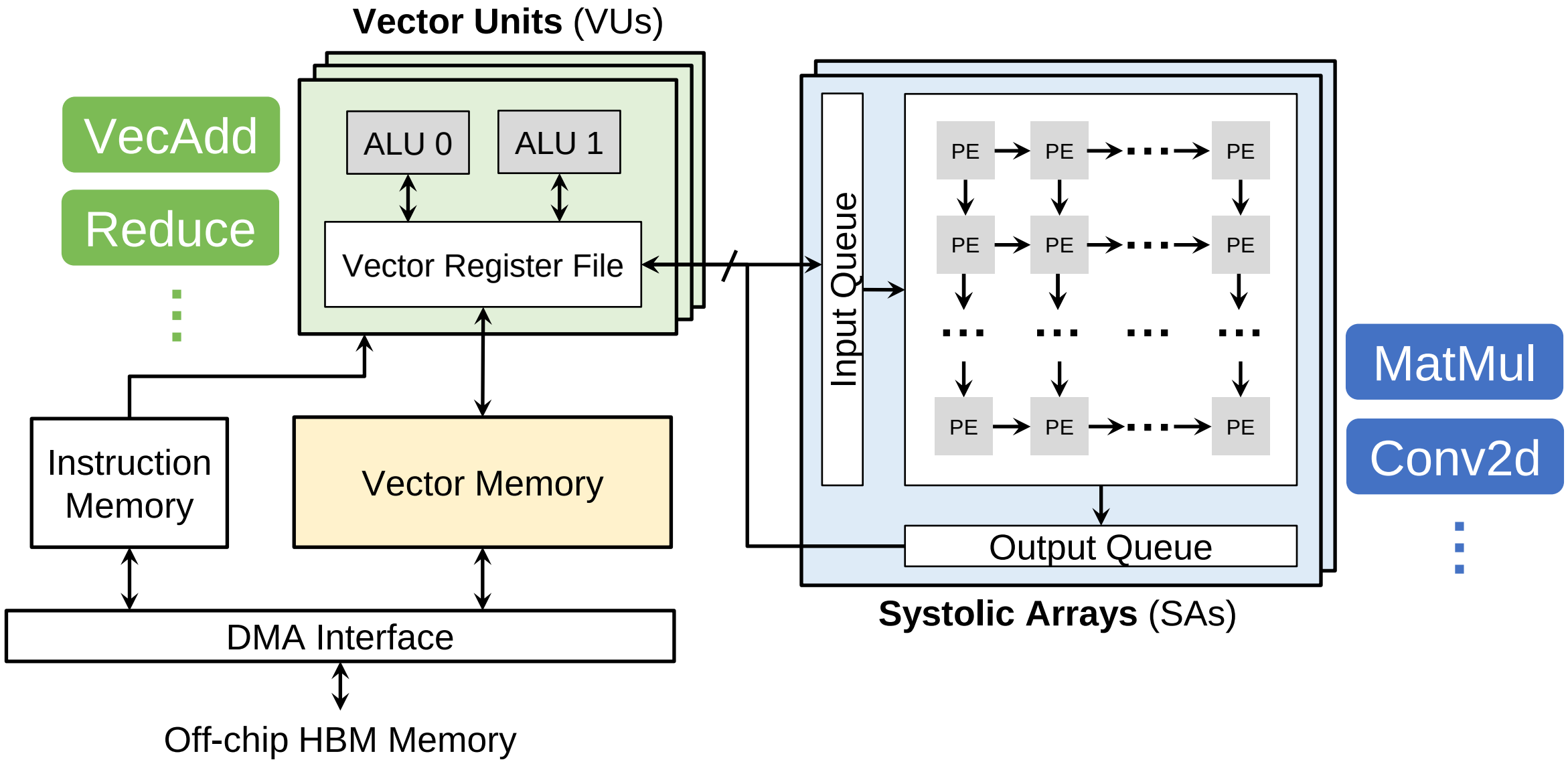
Neural Processing Unit 101



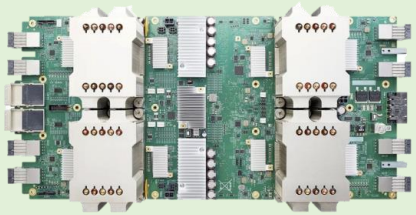
Neural Processing Unit 101



Neural Processing Unit 101



How NPUs Are Used in the Cloud Today?



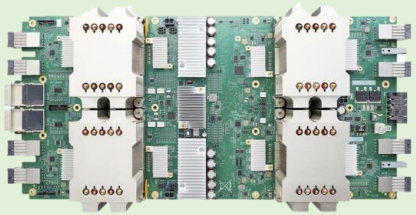
Google TPUv2

How NPUs Are Used in the Cloud Today?

DNN Inference Task
(with Different Batch Sizes)

TensorFlow

TPU VM



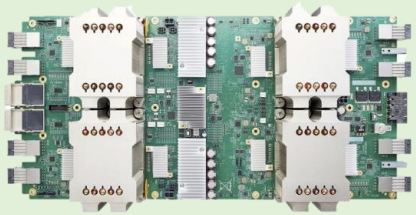
Google TPUv2

How NPUs Are Used in the Cloud Today?

DNN Inference Task
(with Different Batch Sizes)

TensorFlow

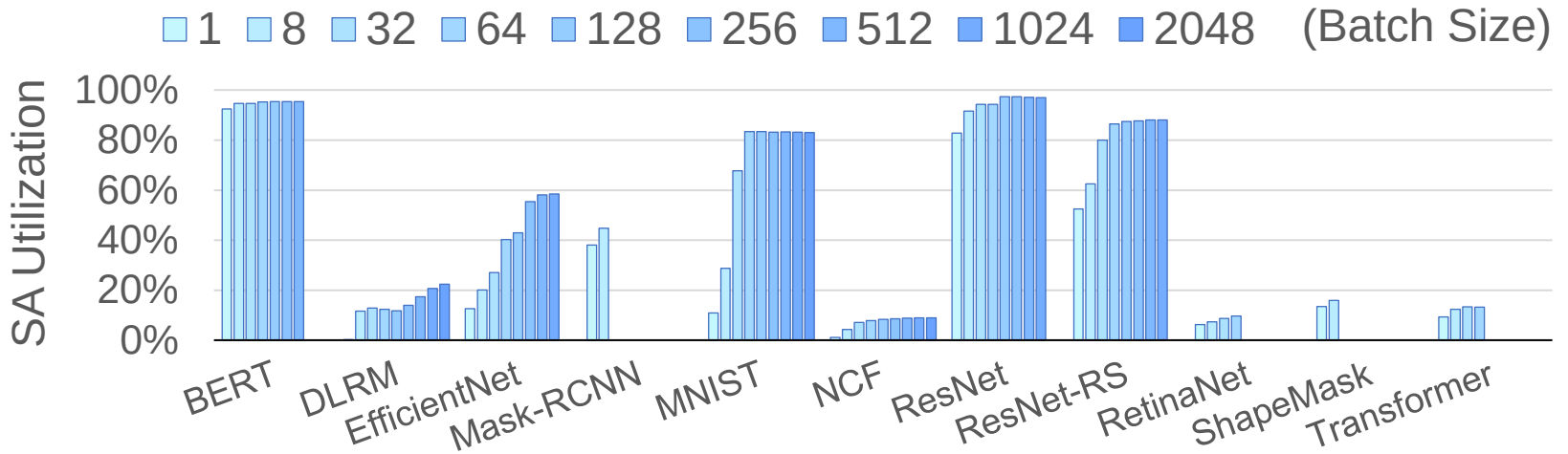
TPU VM



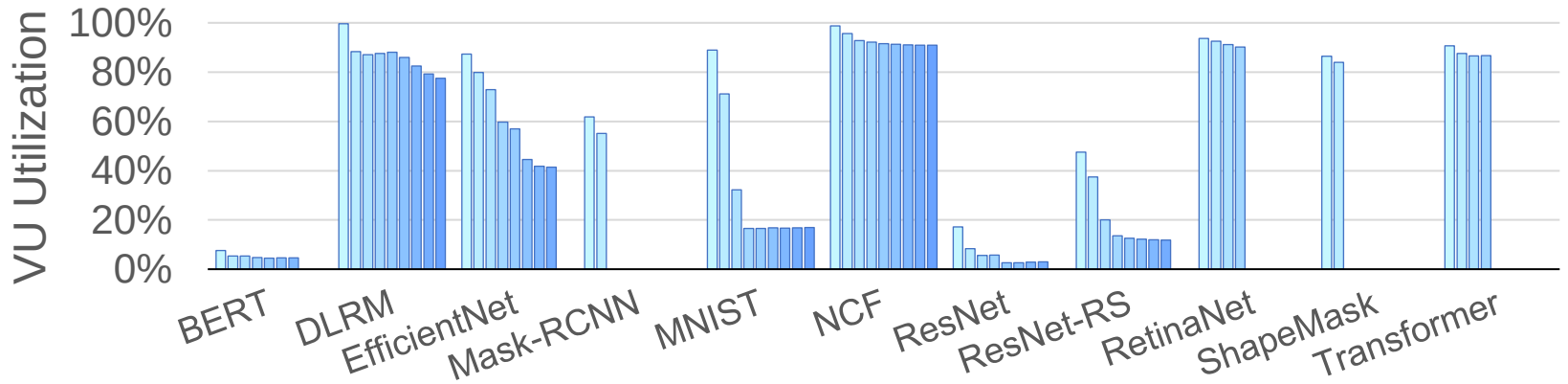
Google TPUv2

Natural Language Processing	BERT, Transformer
Image Classification	ResNet-50, ResNet-RS, EfficientNet, MNIST
Object Detection	ShapeMask, Mask-RCNN, RetinaNet
Recommendation	DLRM, NCF

NPU Cores Are Underutilized When Serving ML Inference Services

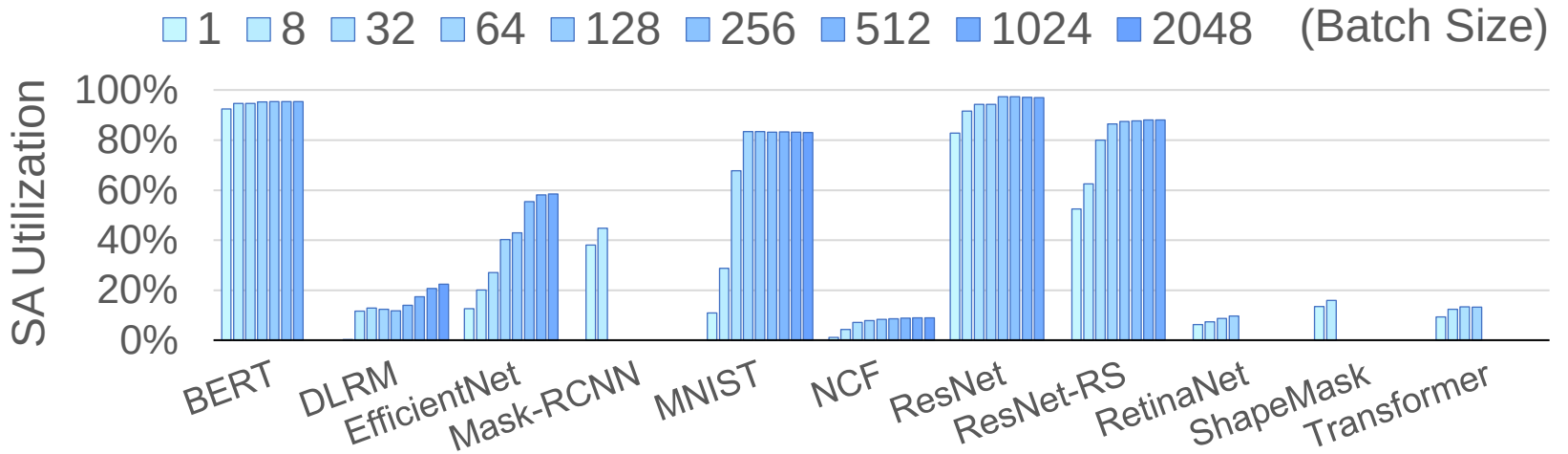


(a) Utilization of Systolic Array

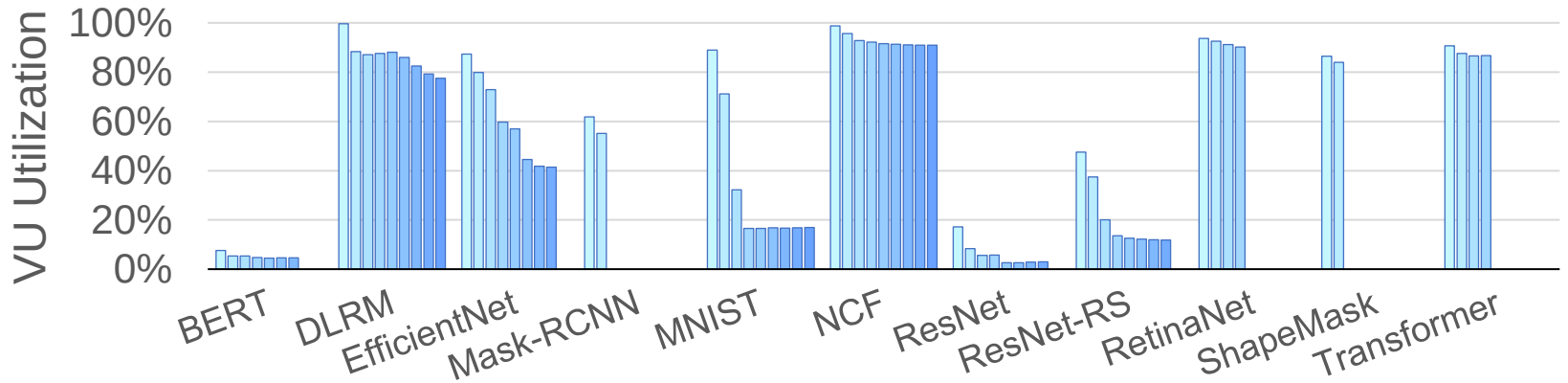


(b) Utilization of Vector Unit

NPU Cores Are Underutilized When Serving ML Inference Services



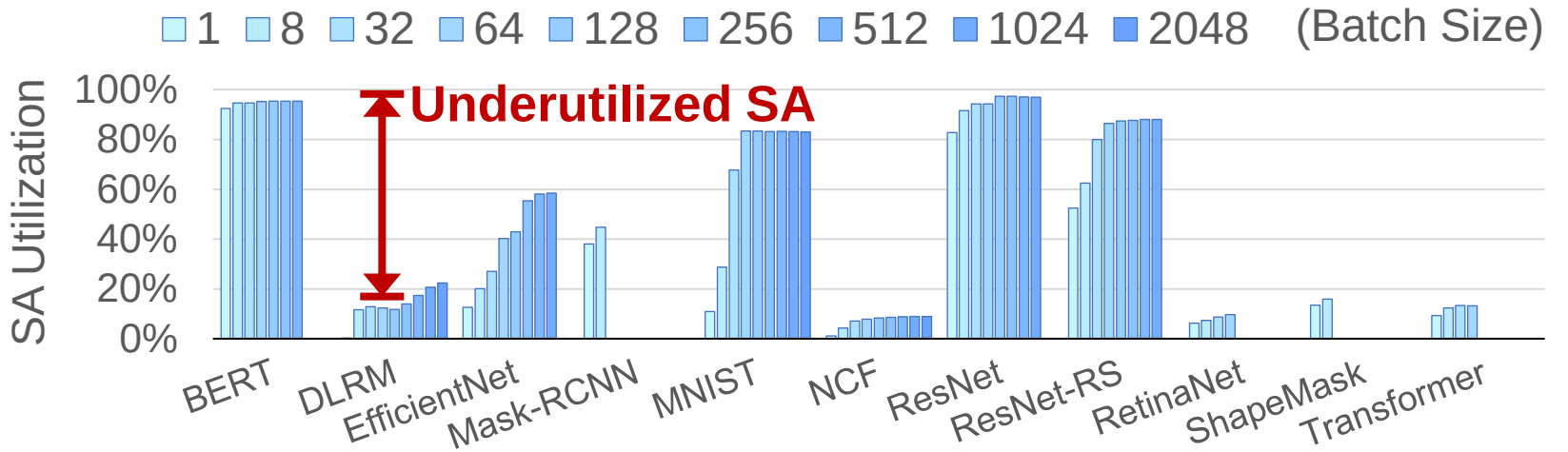
(a) Utilization of Systolic Array



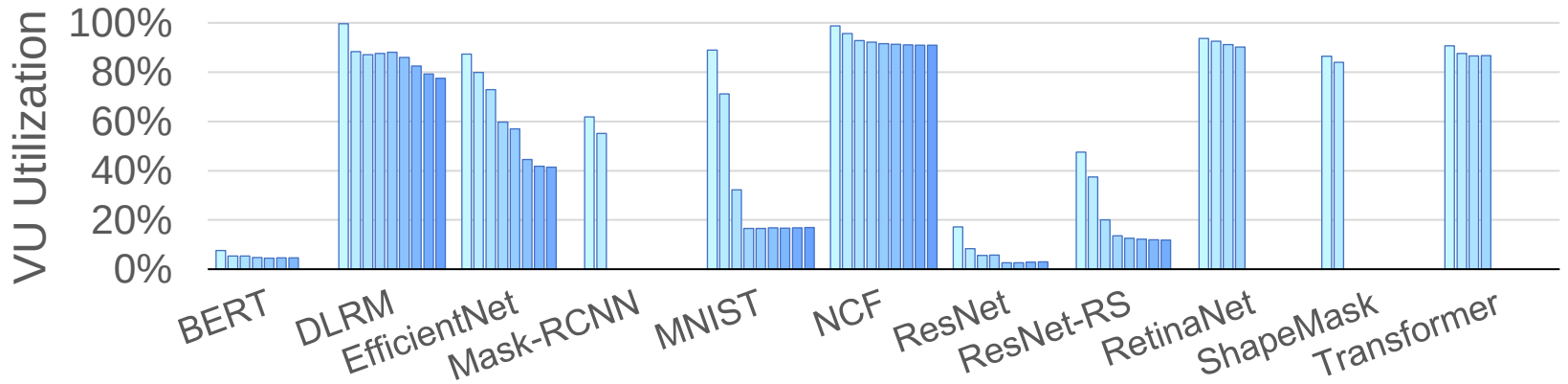
(b) Utilization of Vector Unit

1 Most DNN Inference Workloads Have Imbalanced Use of SAs and VUs

NPU Cores Are Underutilized When Serving ML Inference Services



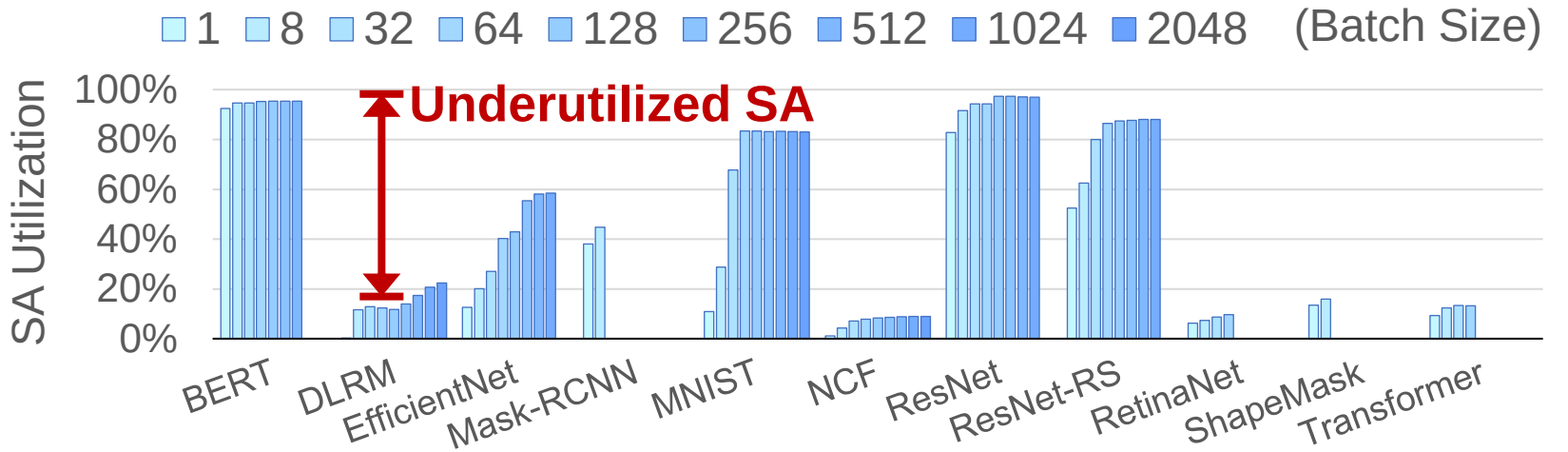
(a) Utilization of Systolic Array



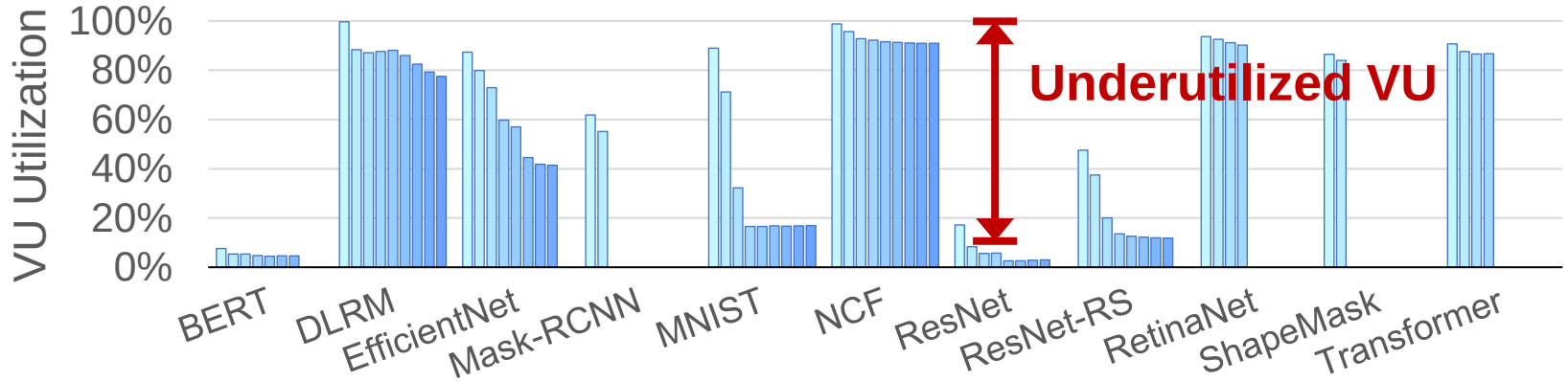
(b) Utilization of Vector Unit

1 Most DNN Inference Workloads Have Imbalanced Use of SAs and VUs

NPU Cores Are Underutilized When Serving ML Inference Services



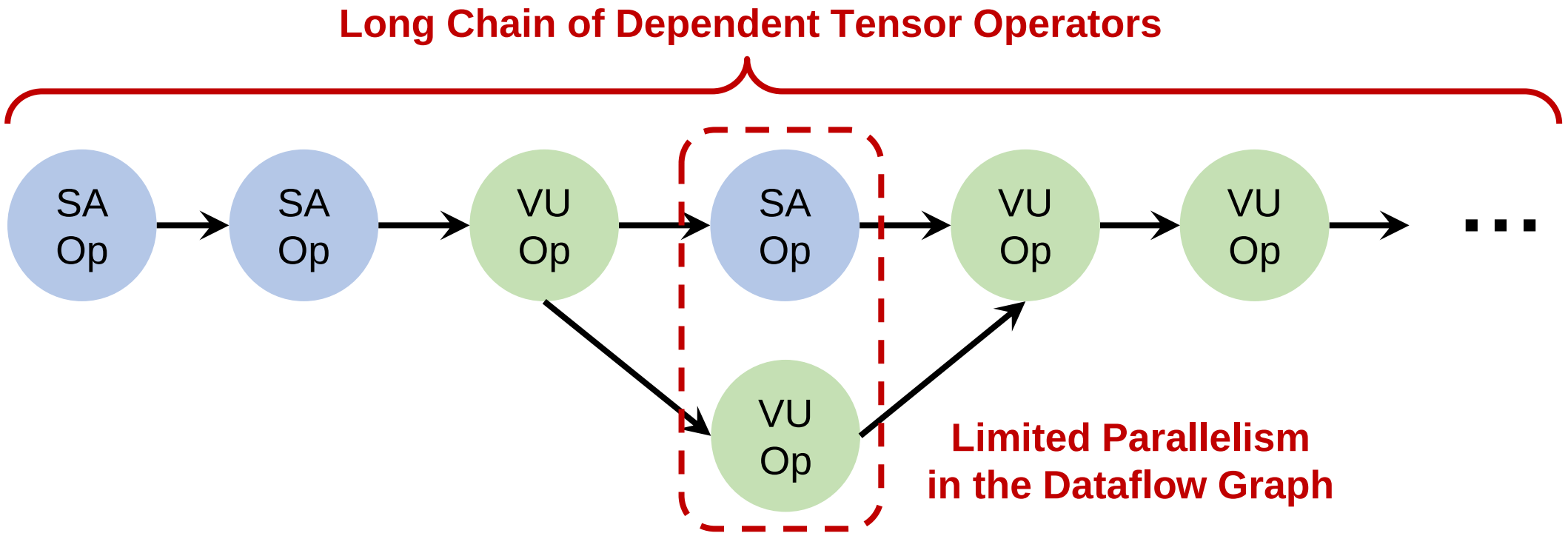
(a) Utilization of Systolic Array



(b) Utilization of Vector Unit

1 Most DNN Inference Workloads Have Imbalanced Use of SAs and VUs

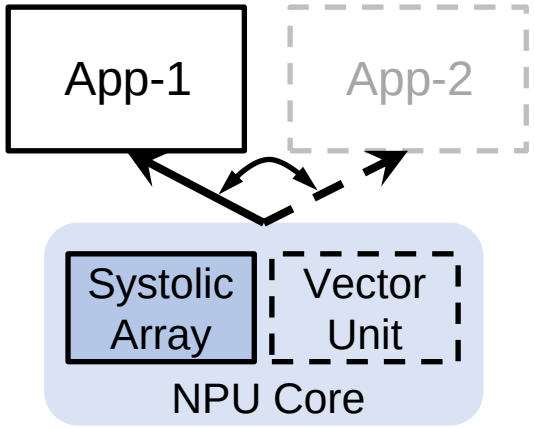
NPU Cores Are Underutilized When Serving ML Inference Services



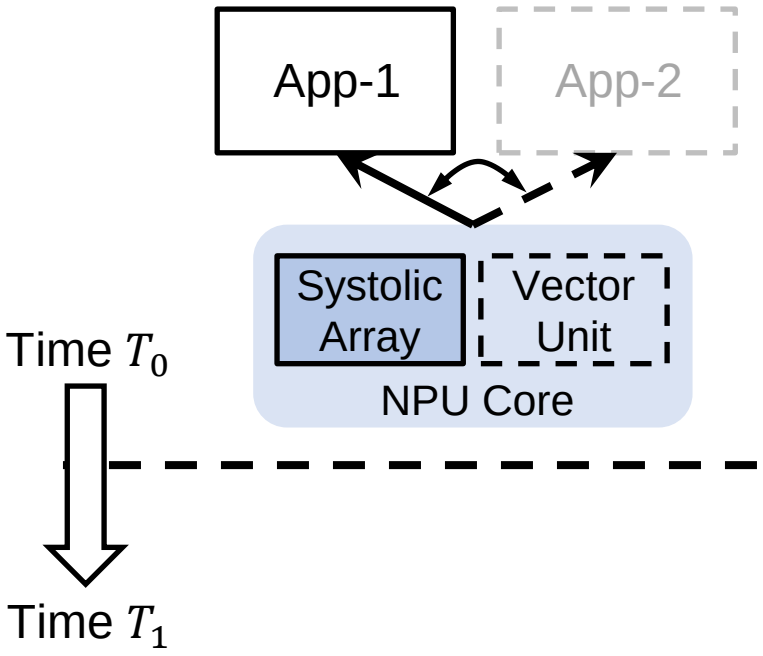
The Dataflow Graph of A Single DNN Workload

2 Most DNN Workloads Have Intensive Data Dependencies Between Operators

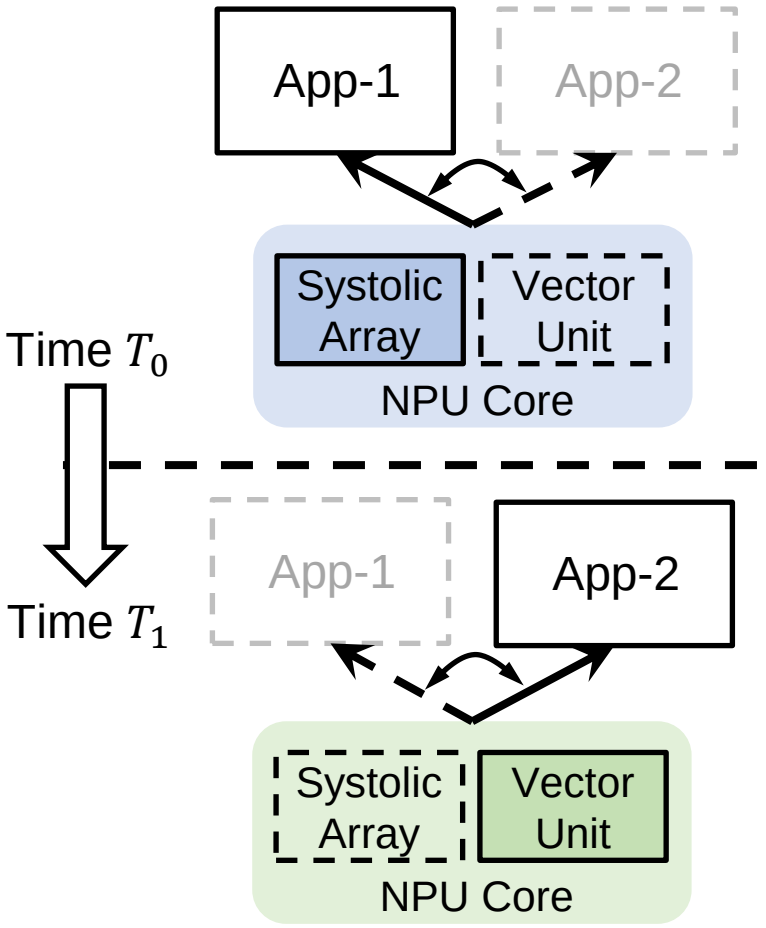
Can NPU Multi-tenancy Today Improve Resource Utilization?



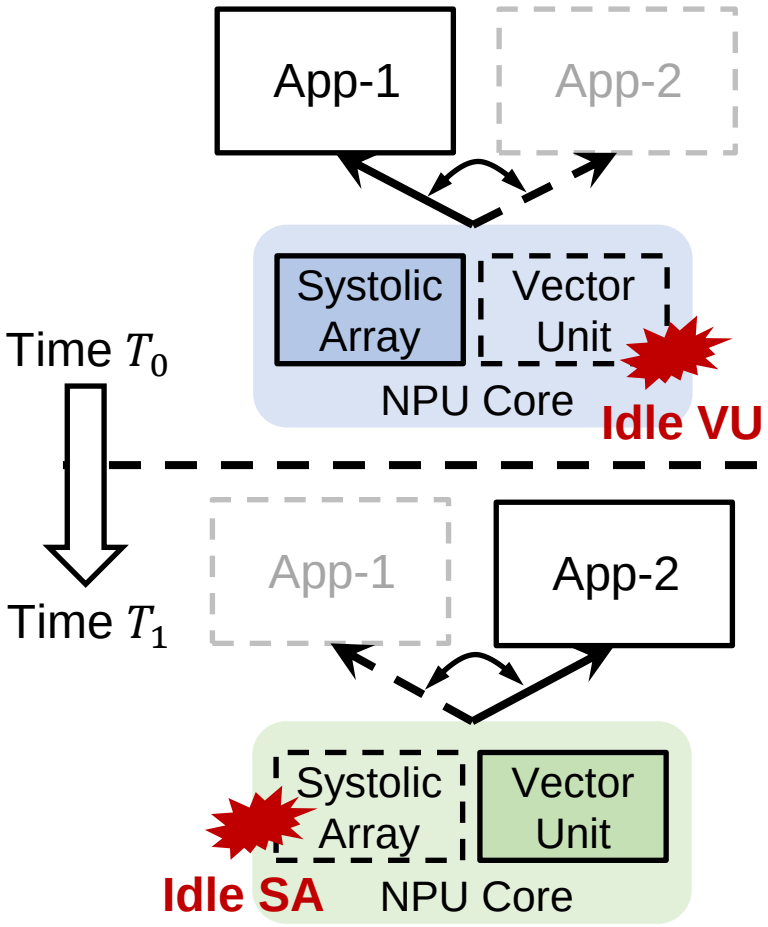
Can NPU Multi-tenancy Today Improve Resource Utilization?



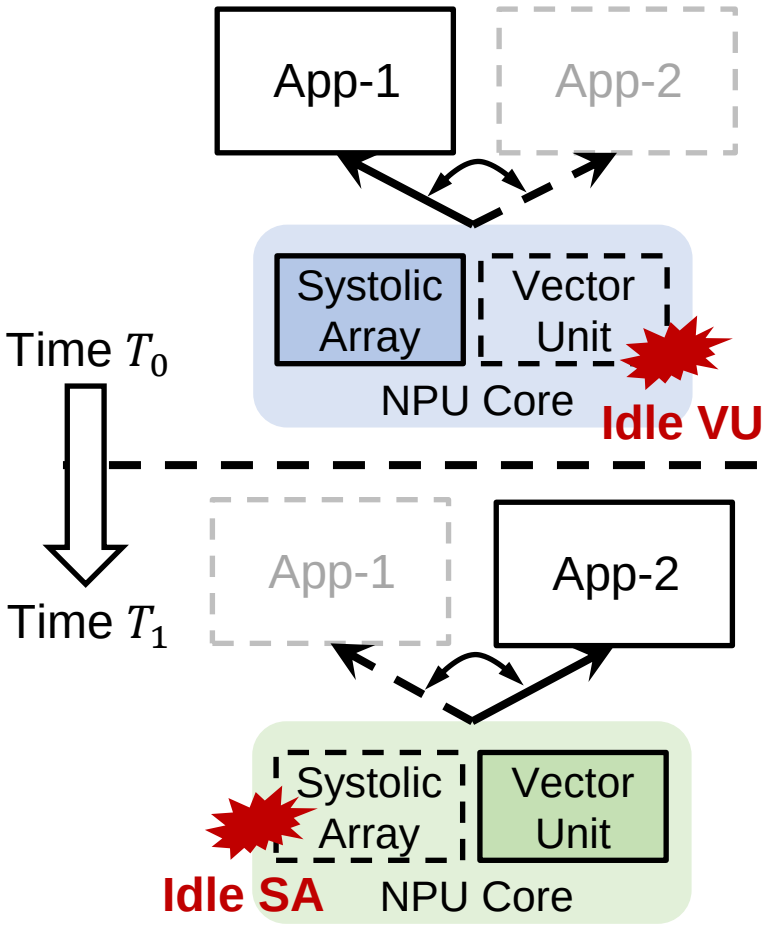
Can NPU Multi-tenancy Today Improve Resource Utilization?



Can NPU Multi-tenancy Today Improve Resource Utilization?

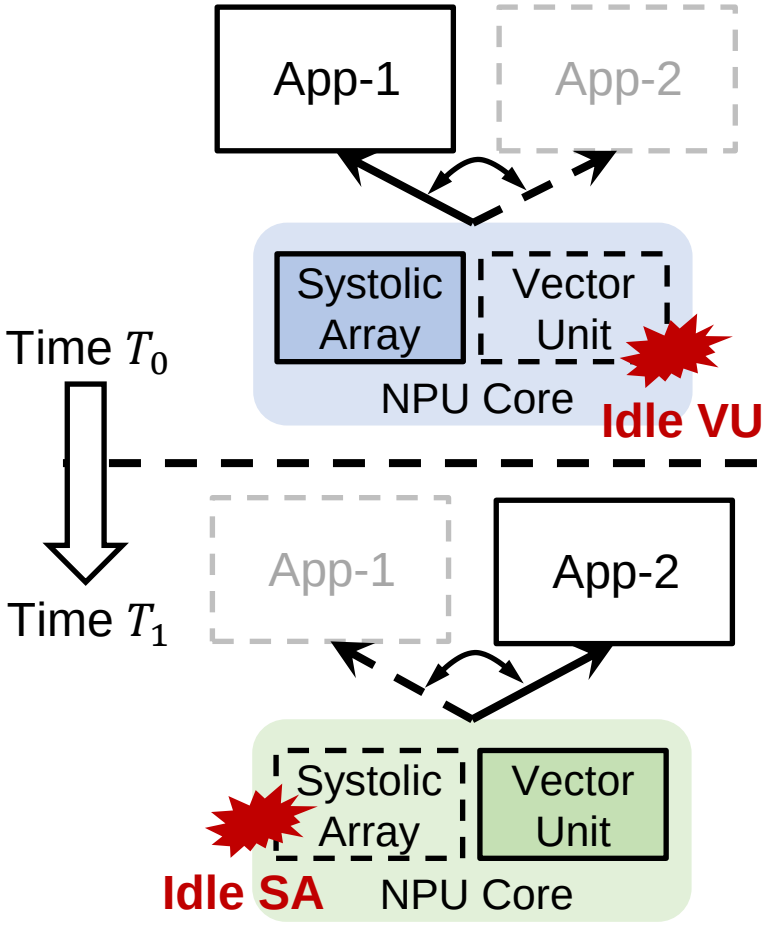


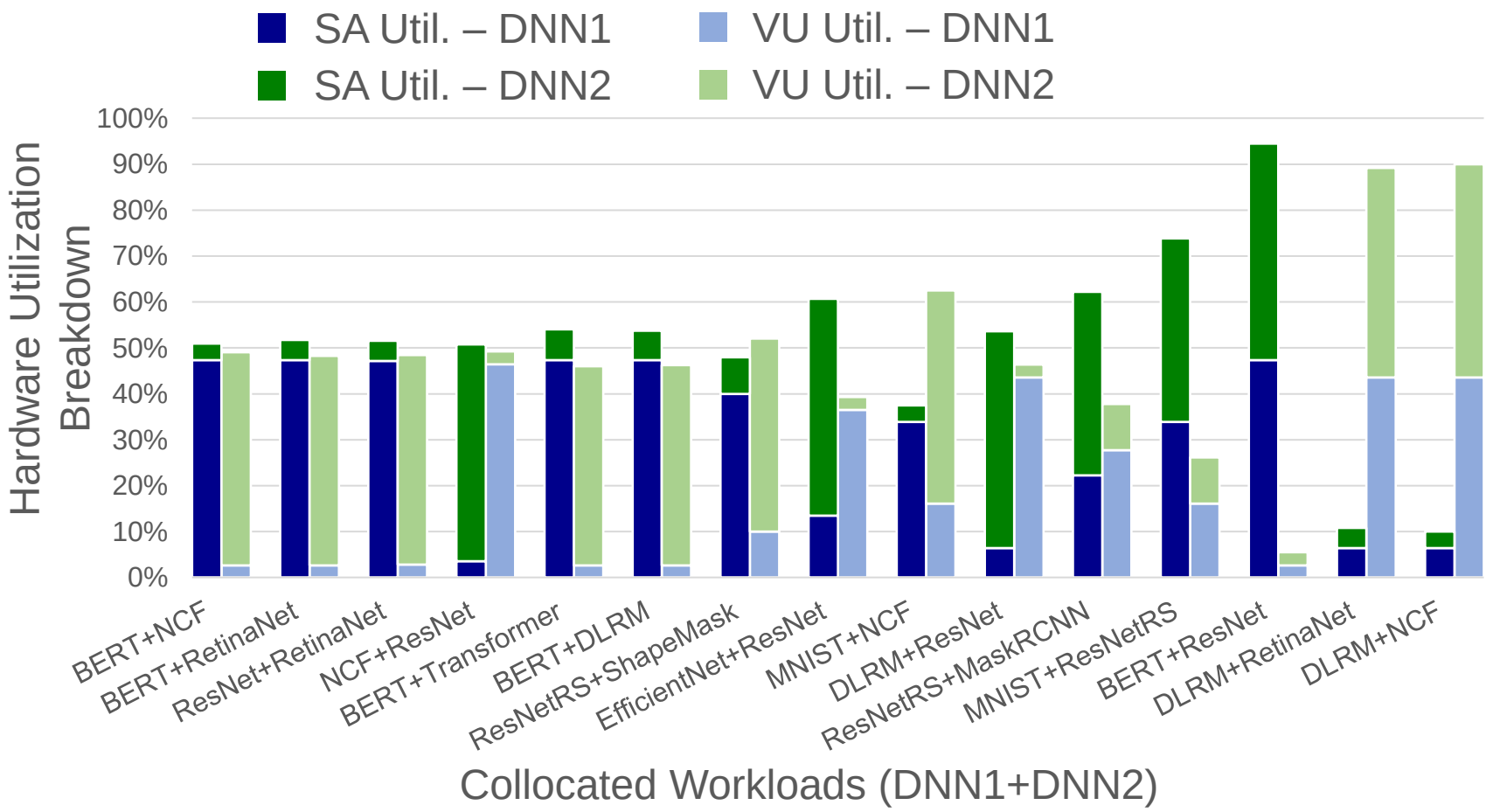
Can NPU Multi-tenancy Today Improve Resource Utilization?



NPU Architecture Today Does Not Support SA/VU-level Resource Sharing

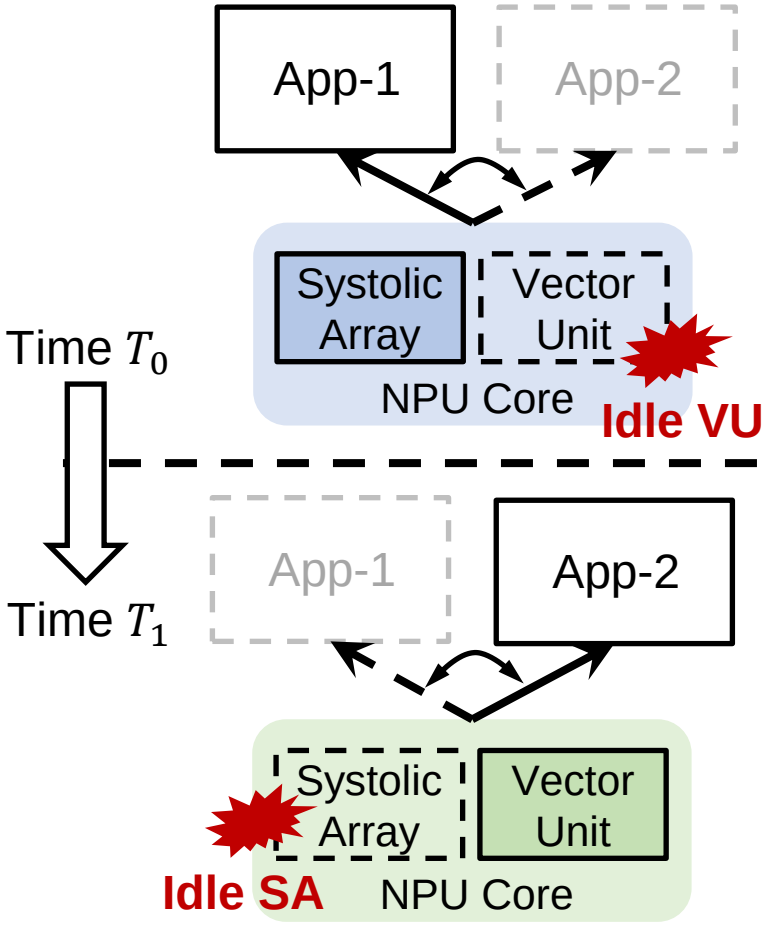
Can NPU Multi-tenancy Today Improve Resource Utilization?

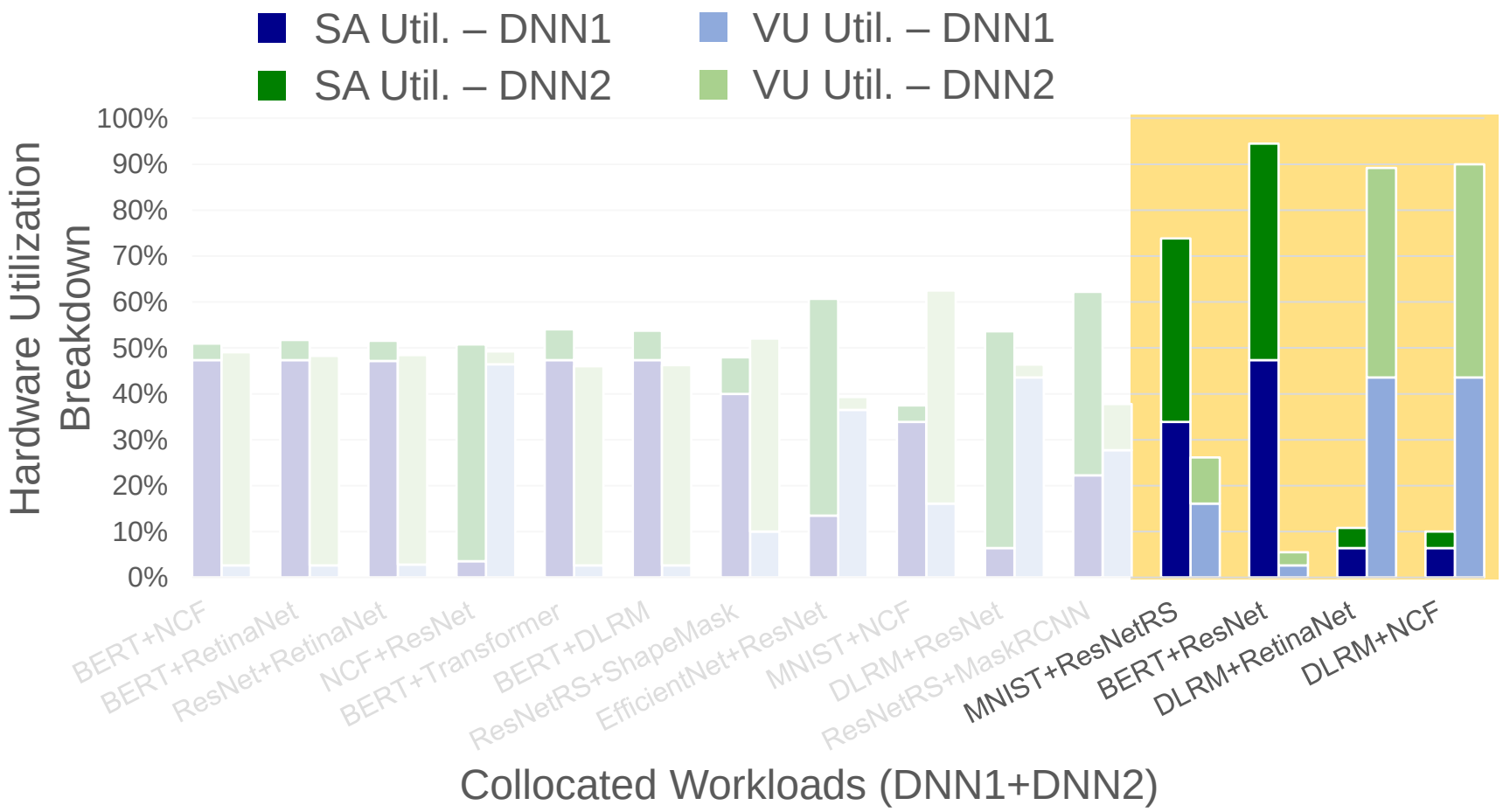




NPU Architecture Today Does Not Support SA/VU-level Resource Sharing

Can NPU Multi-tenancy Today Improve Resource Utilization?

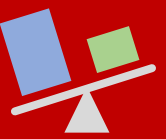




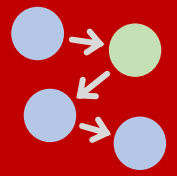
NPU Architecture Today Does Not Support SA/VU-level Resource Sharing

V10: Architectural And System Support for NPU Multi-tenancy

V10: Architectural And System Support for NPU Multi-tenancy

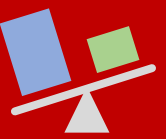


Imbalanced Use of SAs and VUs

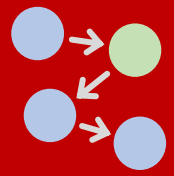


Intensive Data Dependencies in a Single DNN Workload

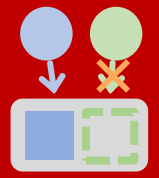
V10: Architectural And System Support for NPU Multi-tenancy



Imbalanced Use of SAs and VUs

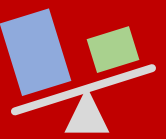


Intensive Data Dependencies in a Single DNN Workload

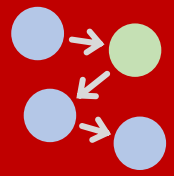


Lack of Architectural Support for NPU Multi-tenancy

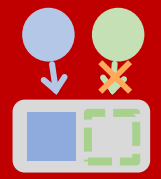
V10: Architectural And System Support for NPU Multi-tenancy



Imbalanced Use of SAs and VUs



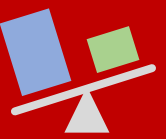
Intensive Data Dependencies in a Single DNN Workload



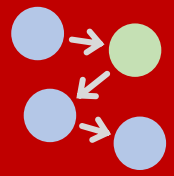
Lack of Architectural Support for NPU Multi-tenancy

V10

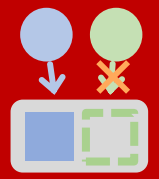
V10: Architectural And System Support for NPU Multi-tenancy



Imbalanced Use of SAs and VUs

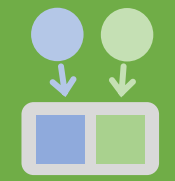


Intensive Data Dependencies in a Single DNN Workload



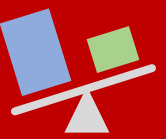
Lack of Architectural Support for NPU Multi-tenancy

HW

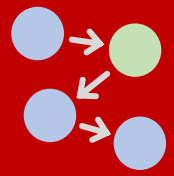


Architectural Support for SA/VU-level Operator Scheduling

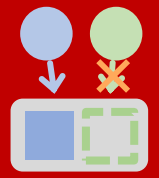
V10: Architectural And System Support for NPU Multi-tenancy



Imbalanced Use of SAs and VUs



Intensive Data Dependencies in a Single DNN Workload

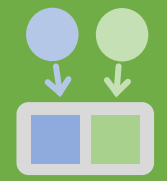


Lack of Architectural Support for NPU Multi-tenancy

SW

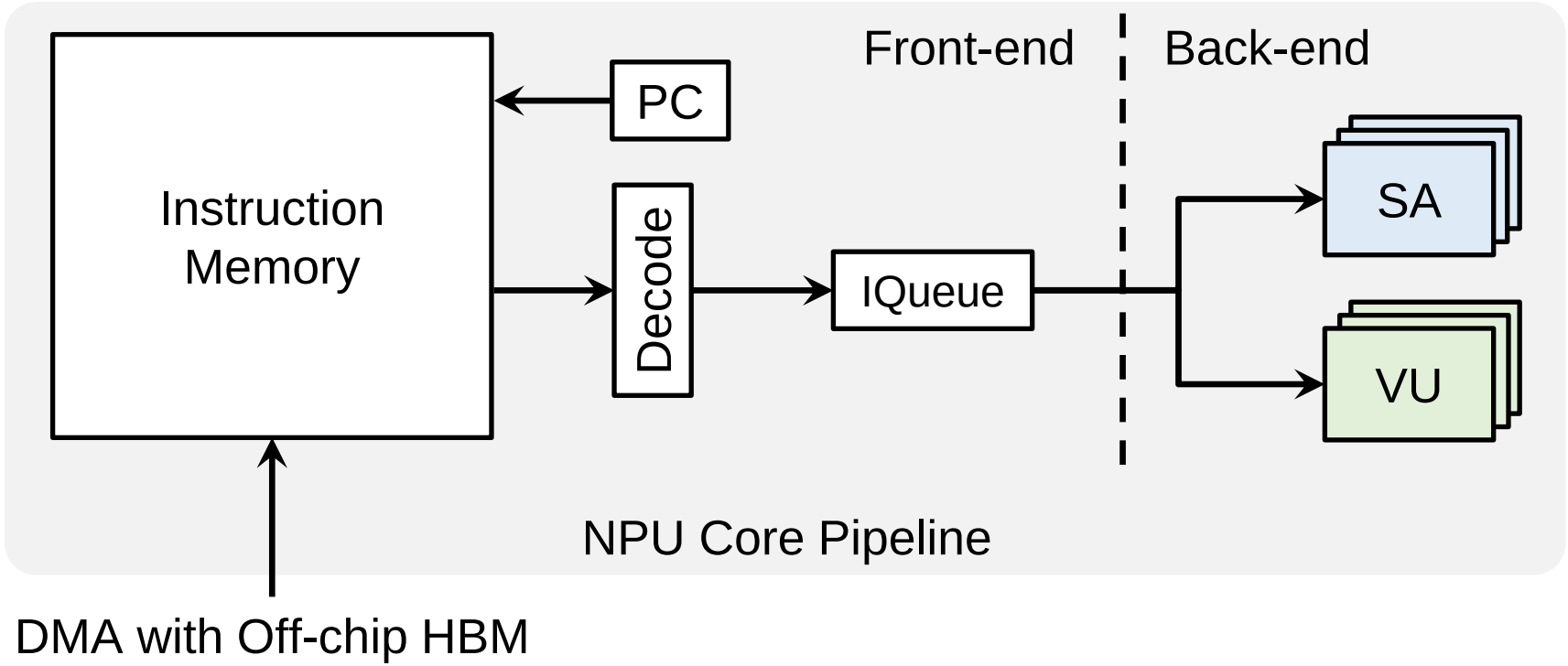
Smart Workload Collocation for Balanced Use of SAs and VUs

HW



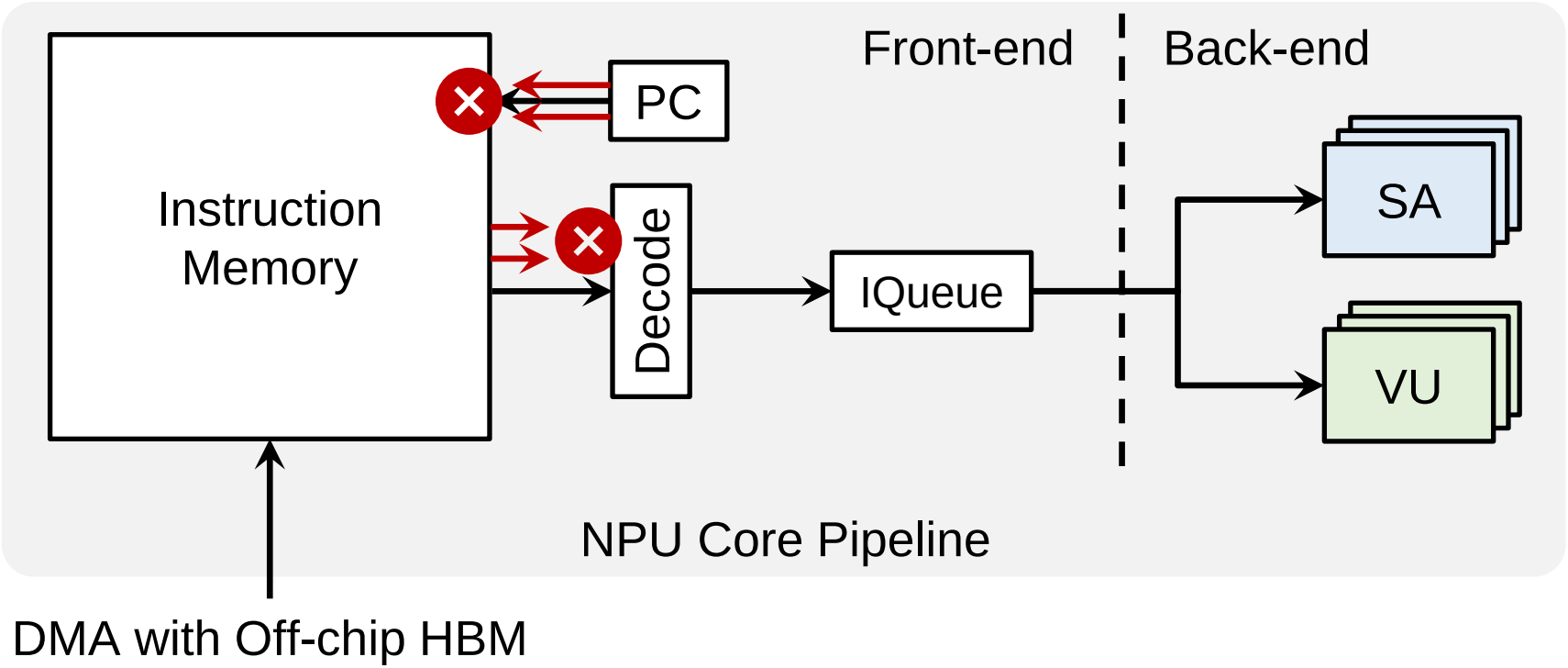
Architectural Support for SA/VU-level Operator Scheduling

Architectural Support for SA/VU-level Operator Scheduling



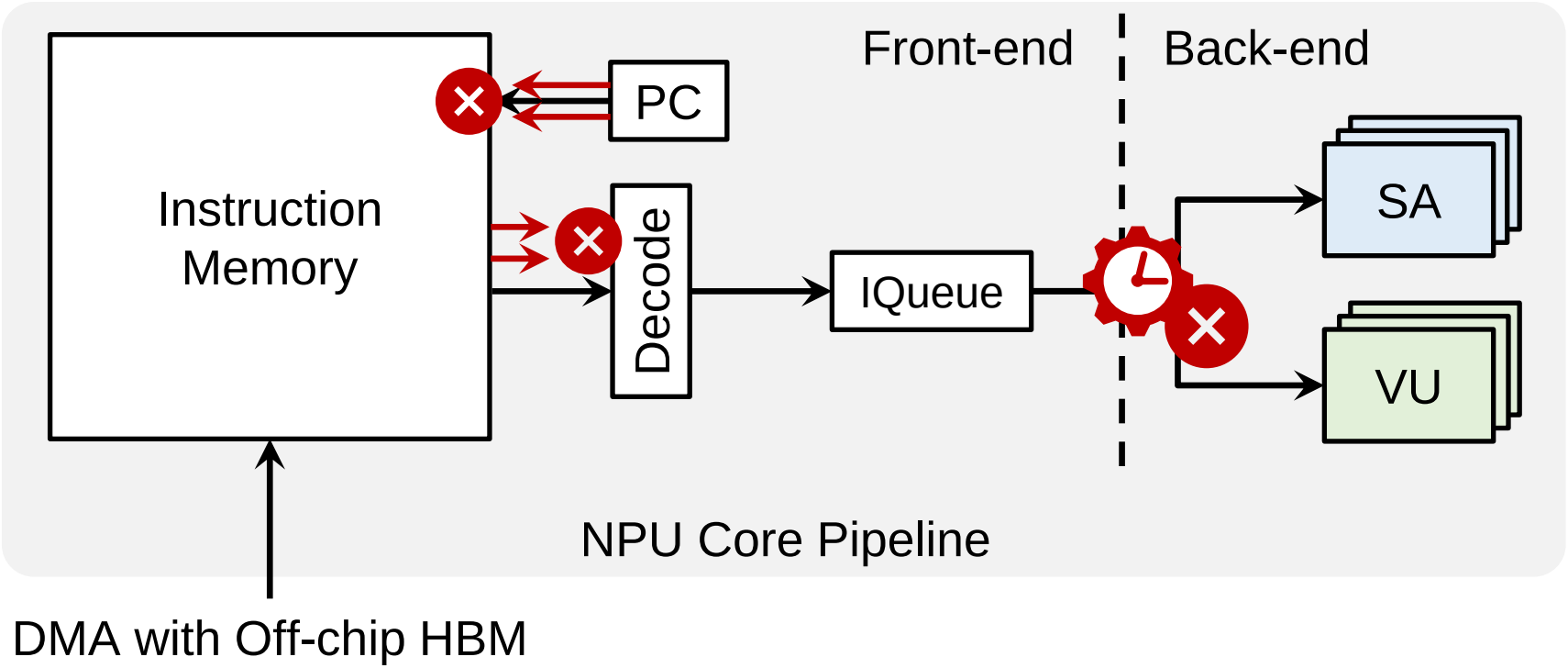
HW

Architectural Support for SA/VU-level Operator Scheduling



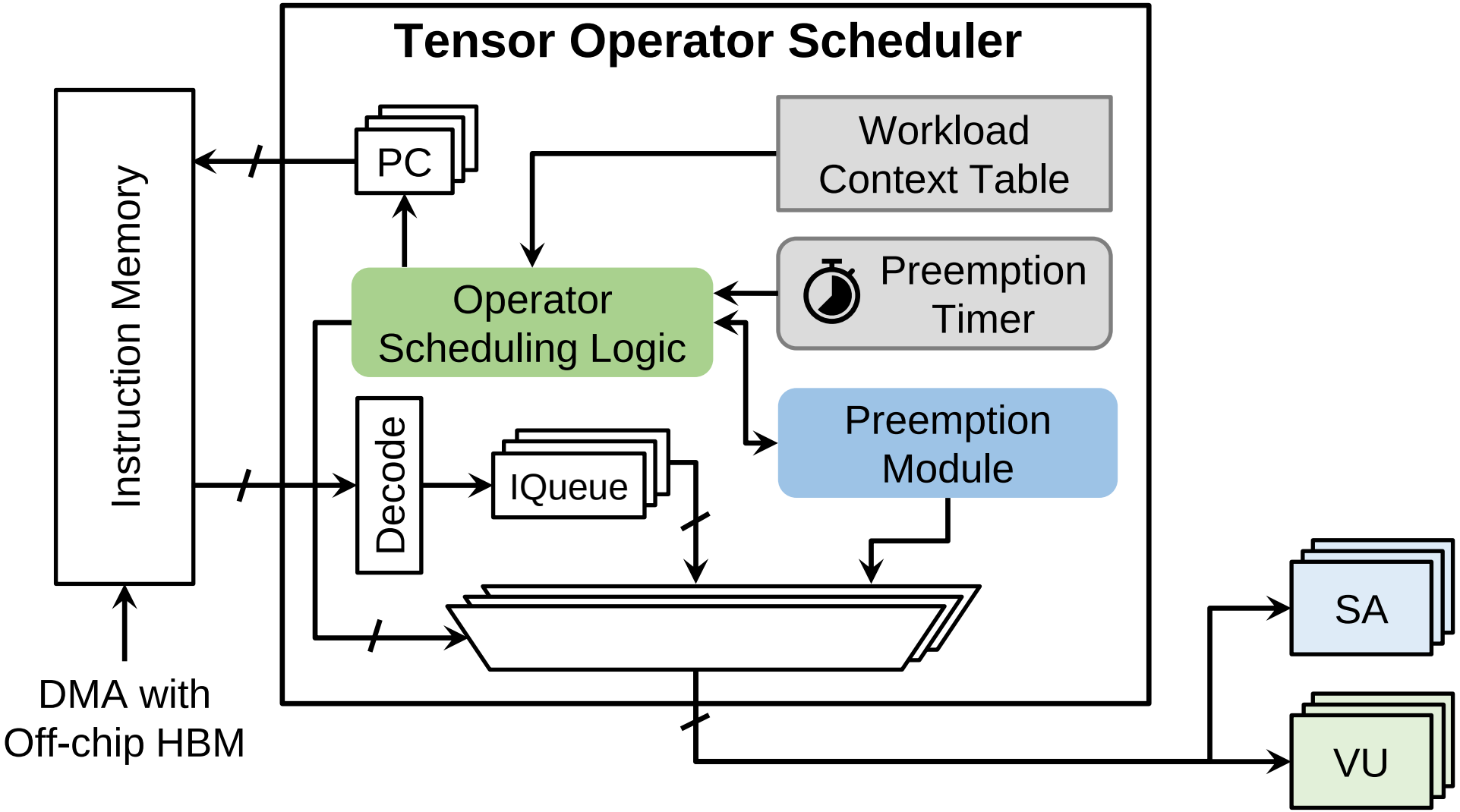
HW

Architectural Support for SA/VU-level Operator Scheduling



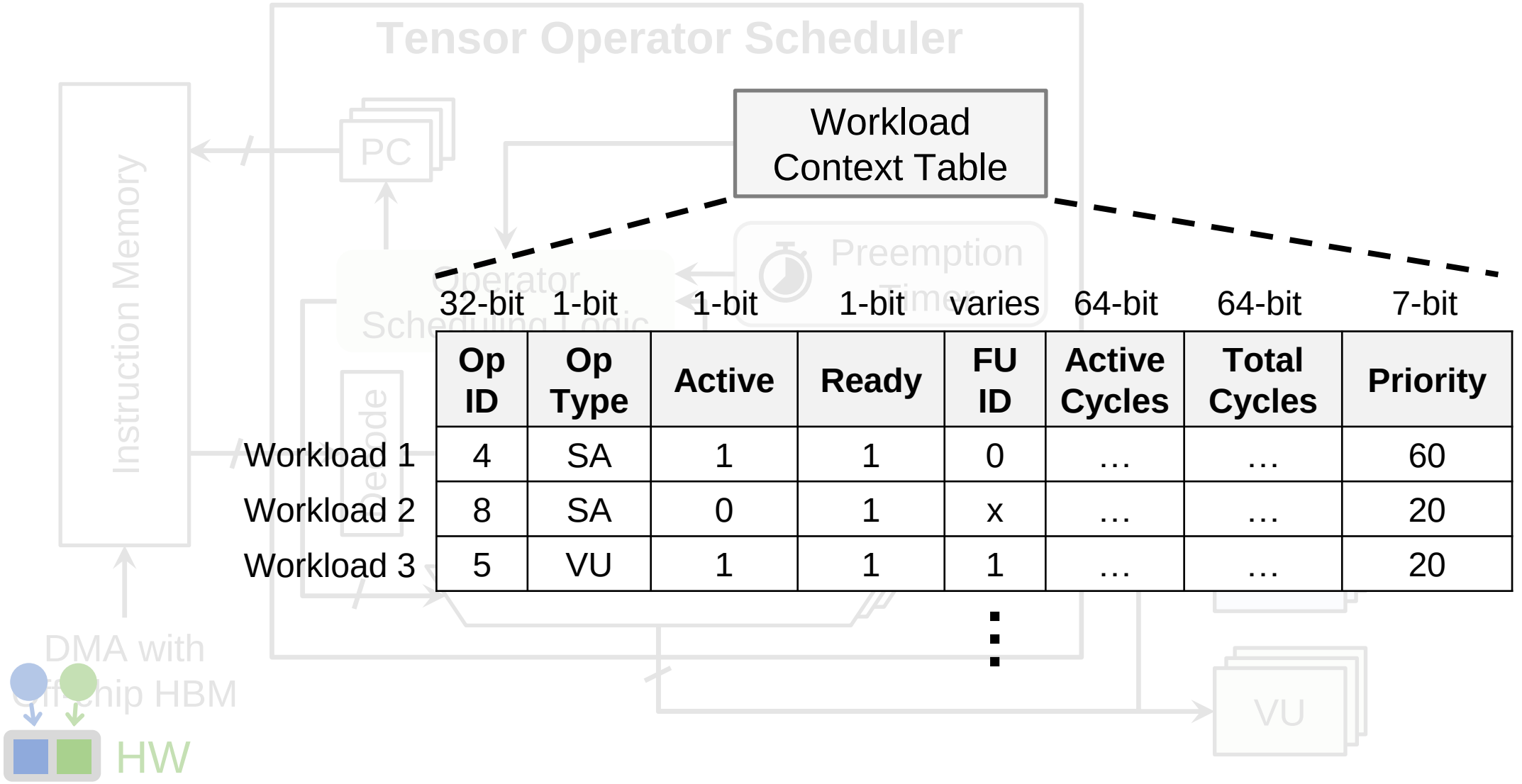
HW

Architectural Support for SA/VU-level Operator Scheduling



HW

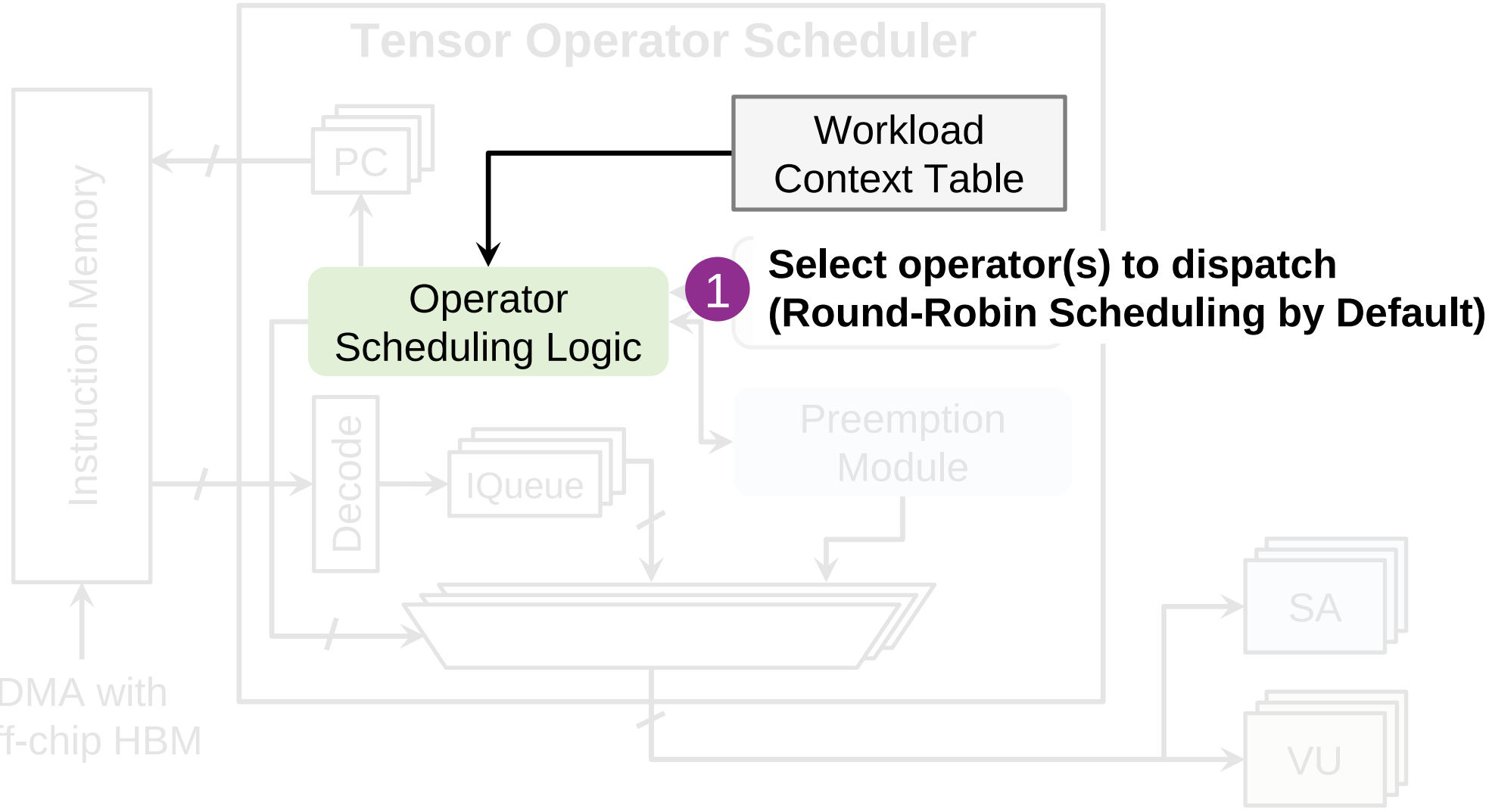
Tracking Multiple DNN Workload Contexts



Workload Context Table

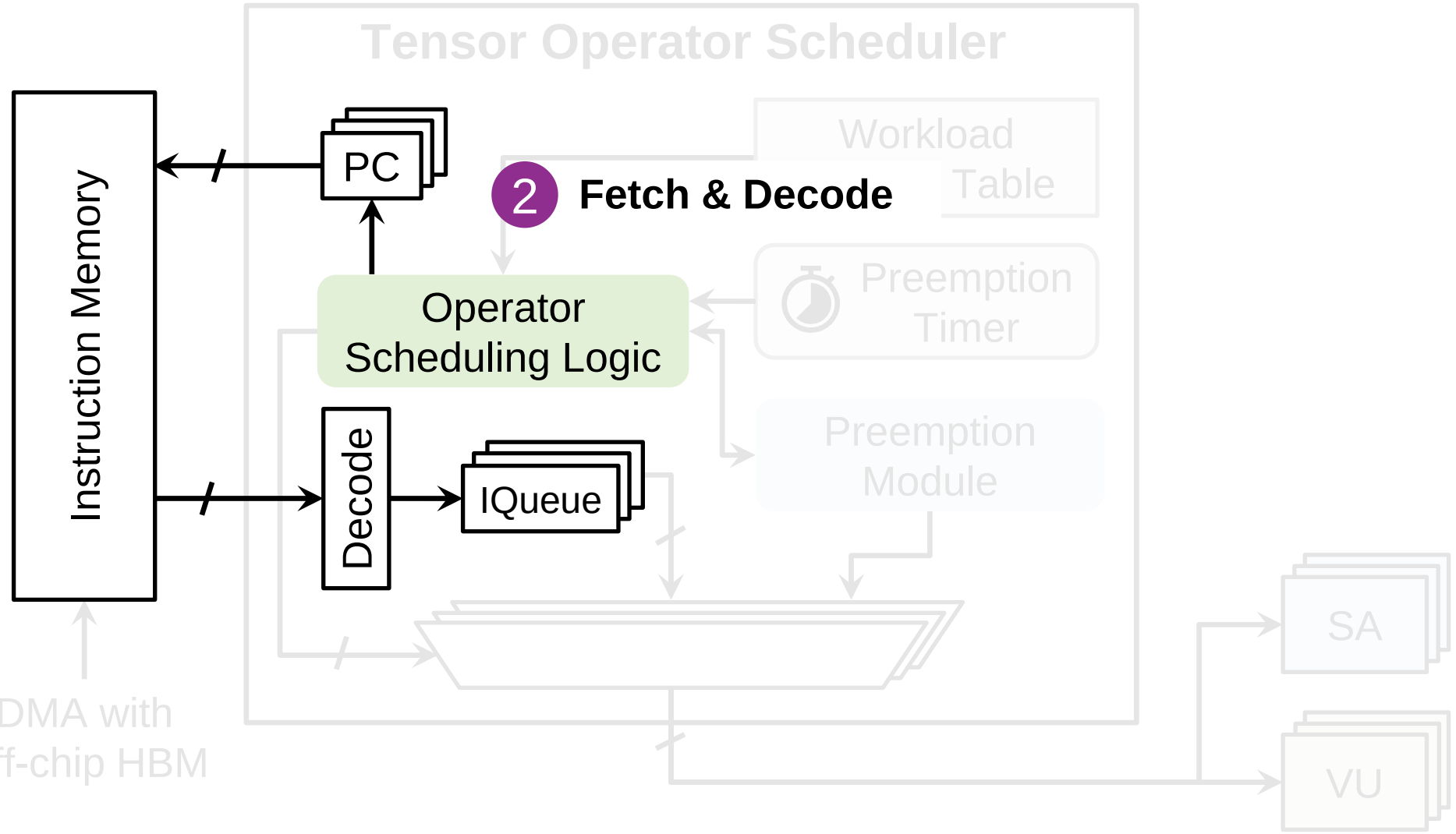
	32-bit	1-bit	1-bit	1-bit	varies	64-bit	64-bit	7-bit
	Op ID	**Op Type**	**Active**	**Ready**	**FU ID**	**Active Cycles**	**Total Cycles**	**Priority**
Workload 1	4	SA	1	1	0	…	…	60
Workload 2	8	SA	0	1	x	…	…	20
Workload 3	5	VU	1	1	1	…	…	20
	⋮							

Dispatching Tensor Operators From Multi-tenant DNN Workloads



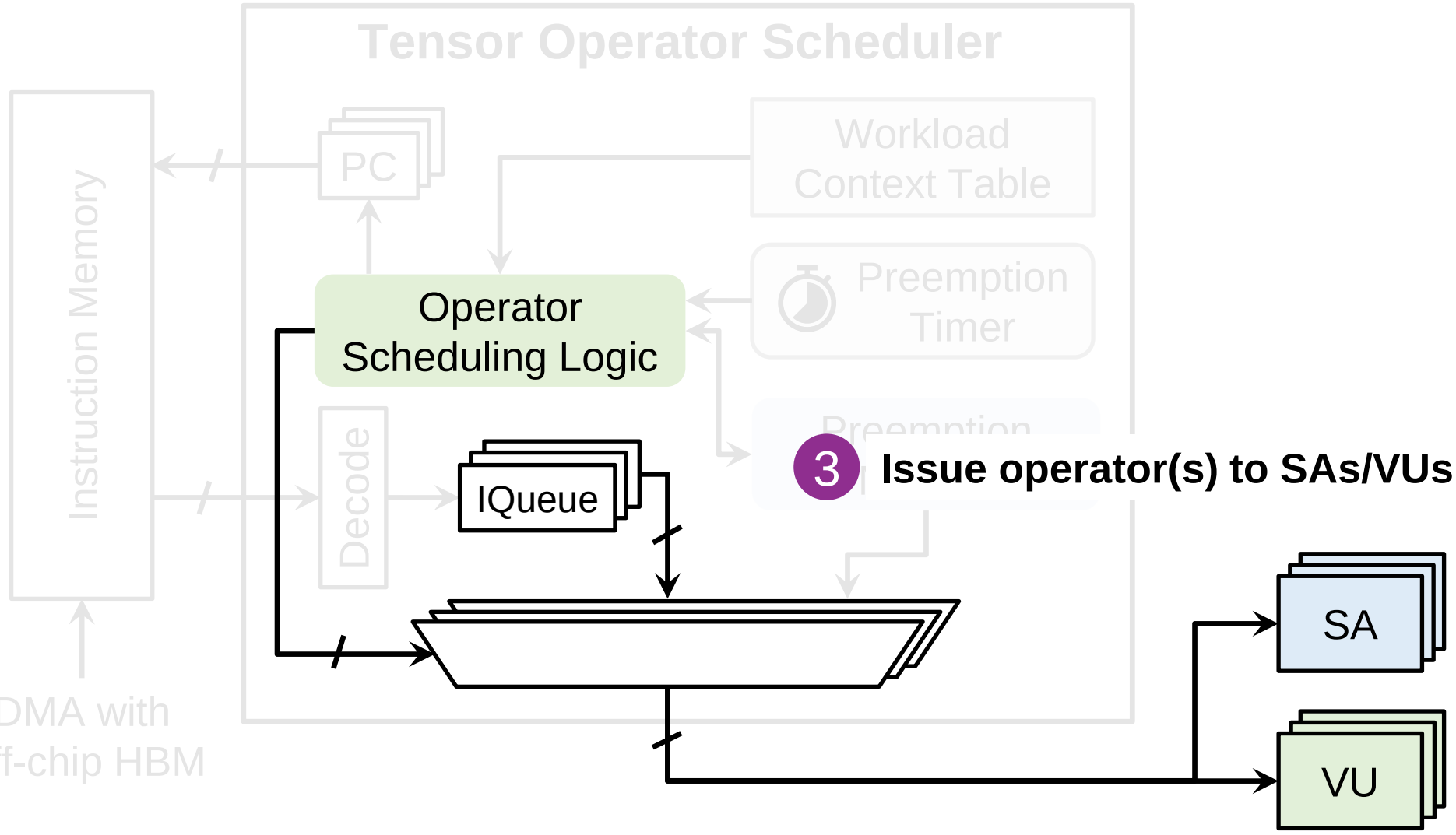
HW

Dispatching Tensor Operators From Multi-tenant DNN Workloads

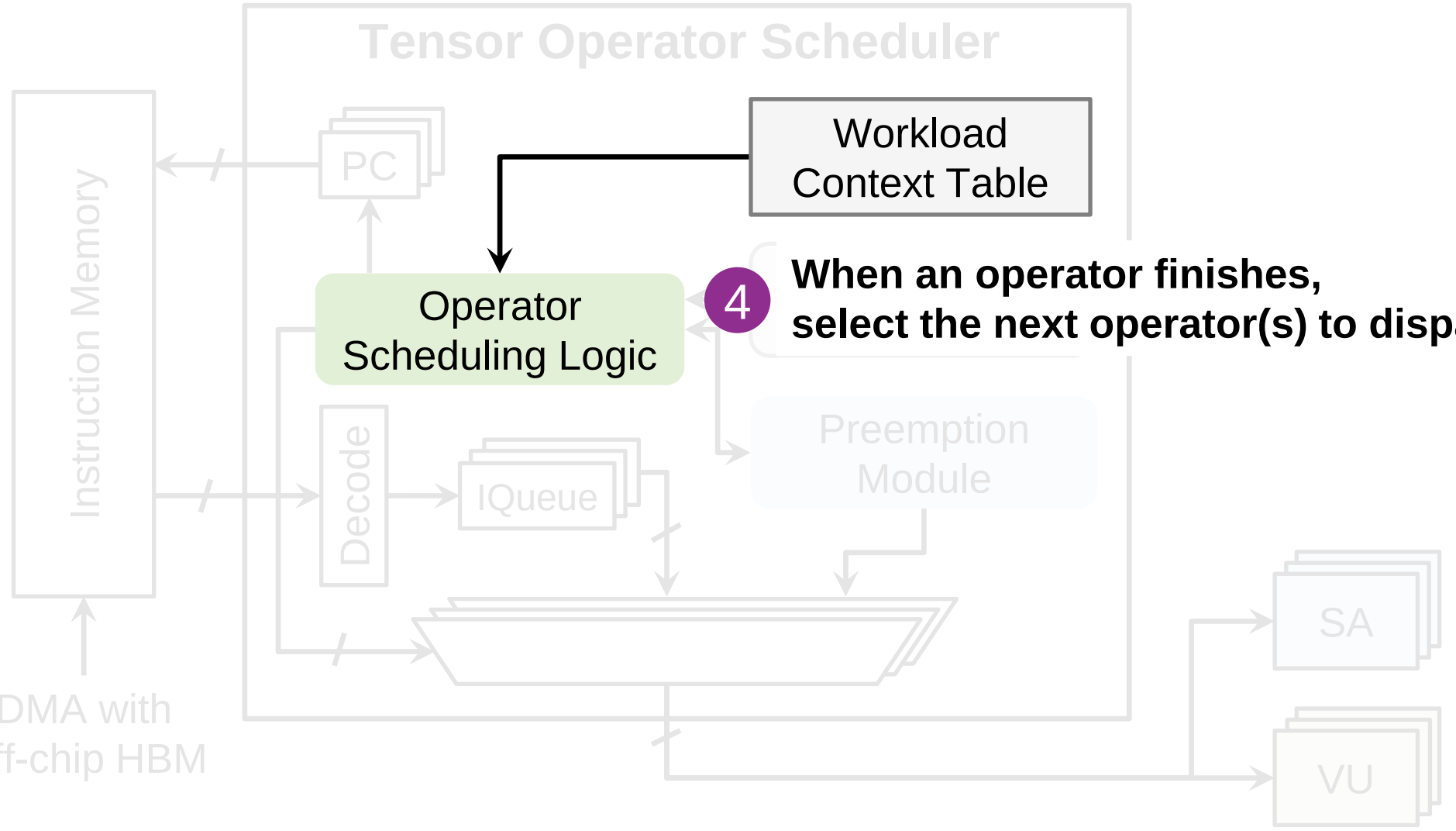


HW

Dispatching Tensor Operators From Multi-tenant DNN Workloads

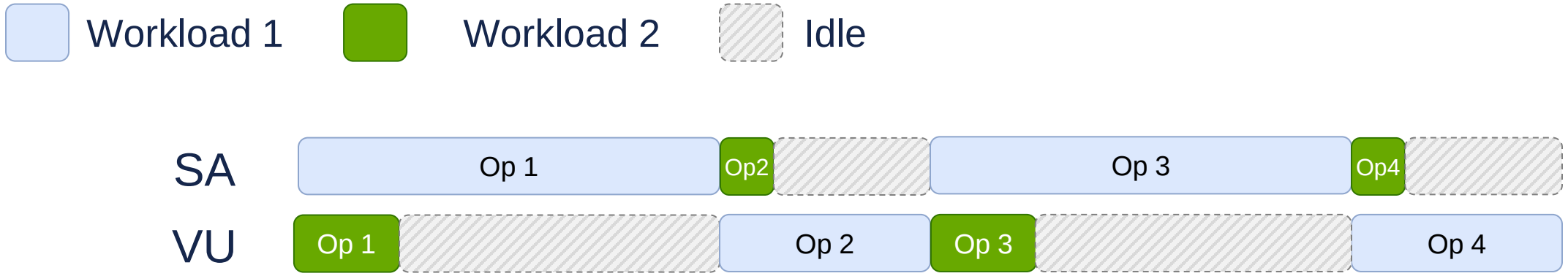


Dispatching Tensor Operators From Multi-tenant DNN Workloads

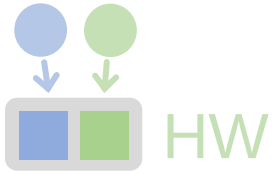


HW

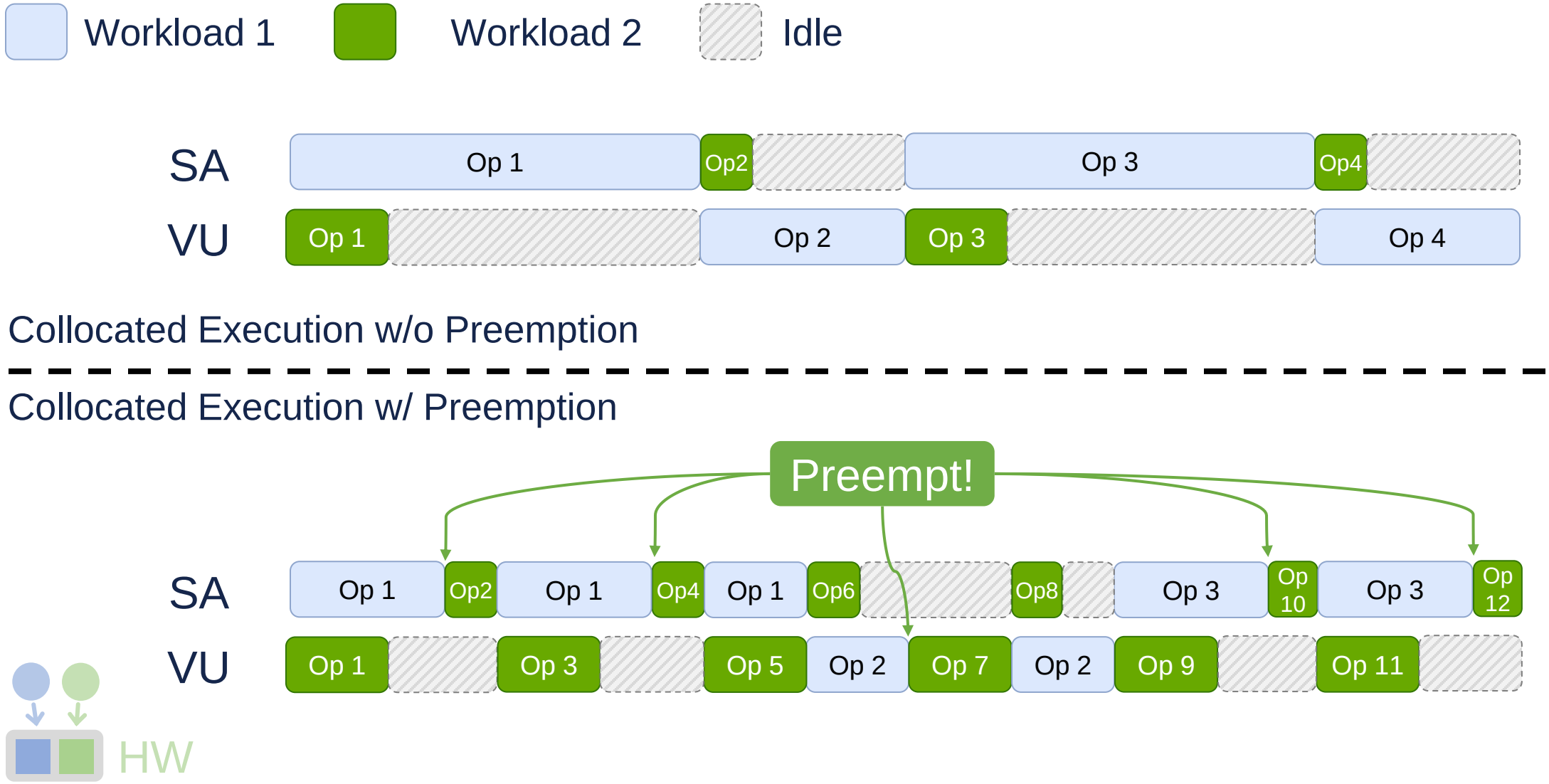
Improving Fairness and Utilization with Tensor Operator Preemption



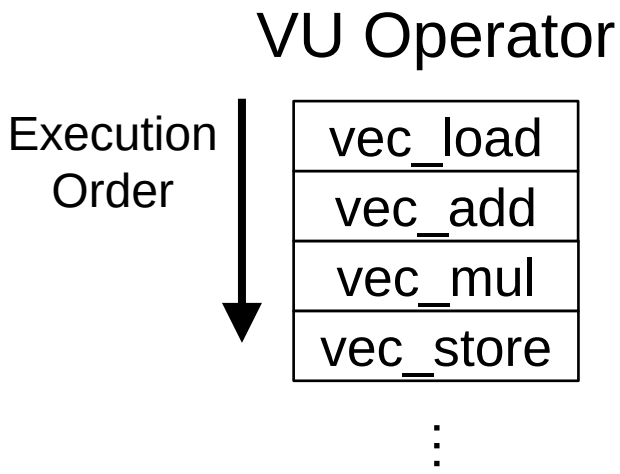
Collocated Execution w/o Preemption

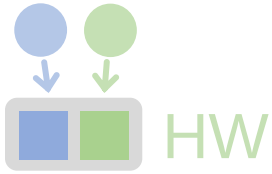


Improving Fairness and Utilization with Tensor Operator Preemption

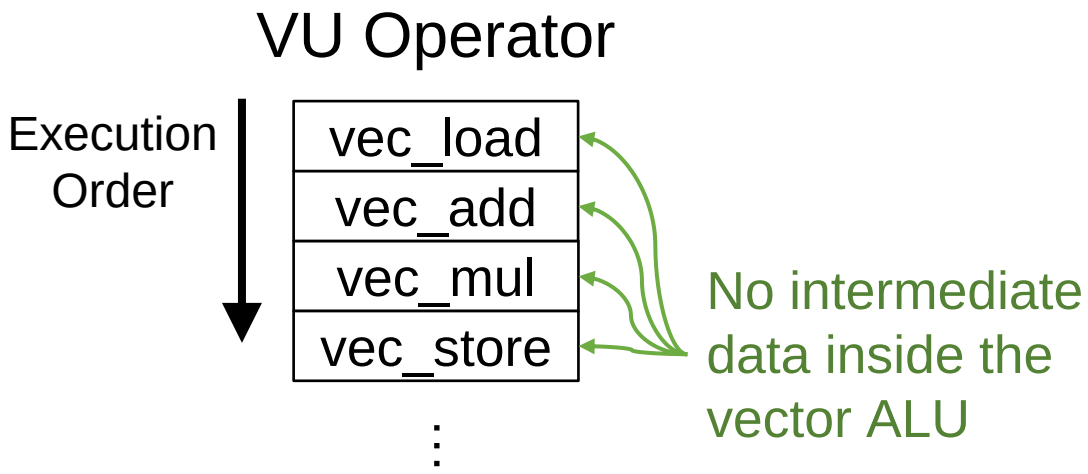


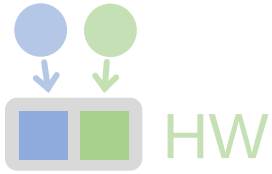
A Lightweight Context Switch Mechanism for Tensor Operator Preemption



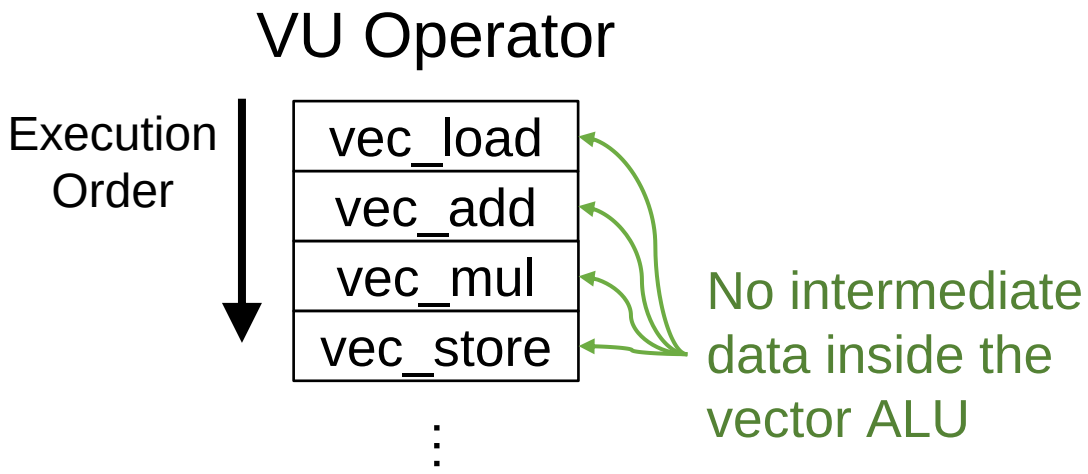


A Lightweight Context Switch Mechanism for Tensor Operator Preemption

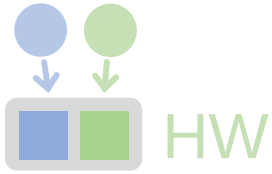




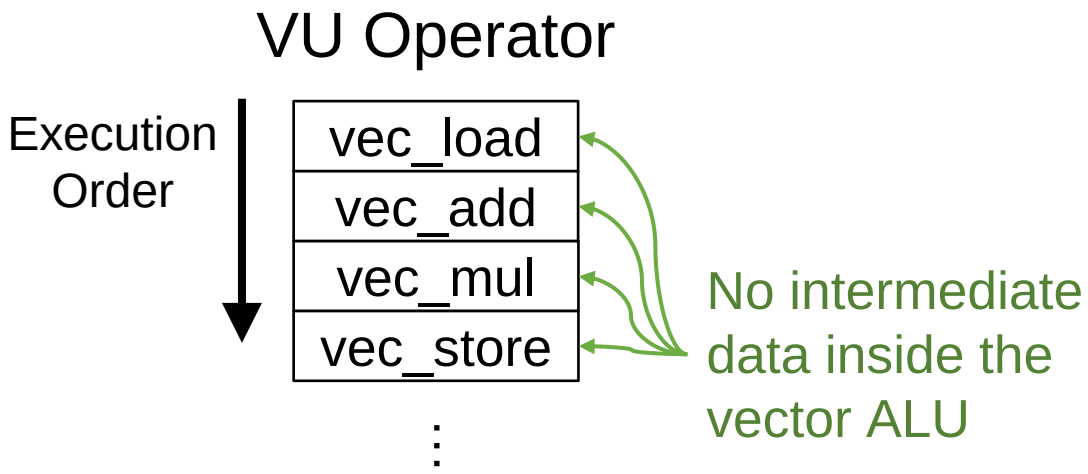
A Lightweight Context Switch Mechanism for Tensor Operator Preemption

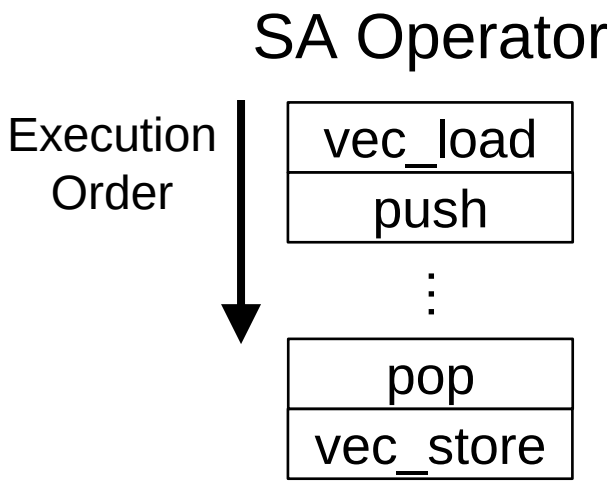


VU Preemption at Instruction Granularity

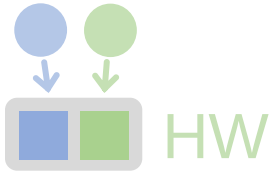


A Lightweight Context Switch Mechanism for Tensor Operator Preemption

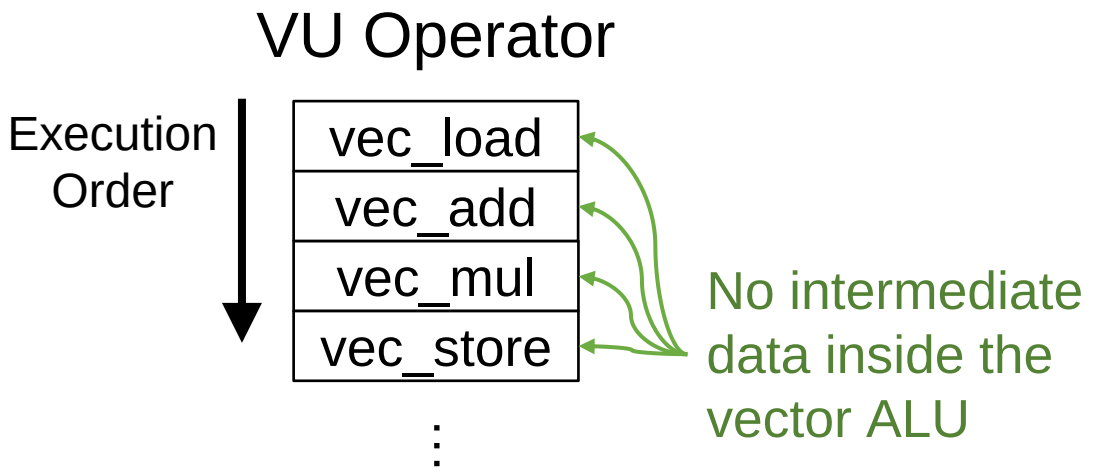


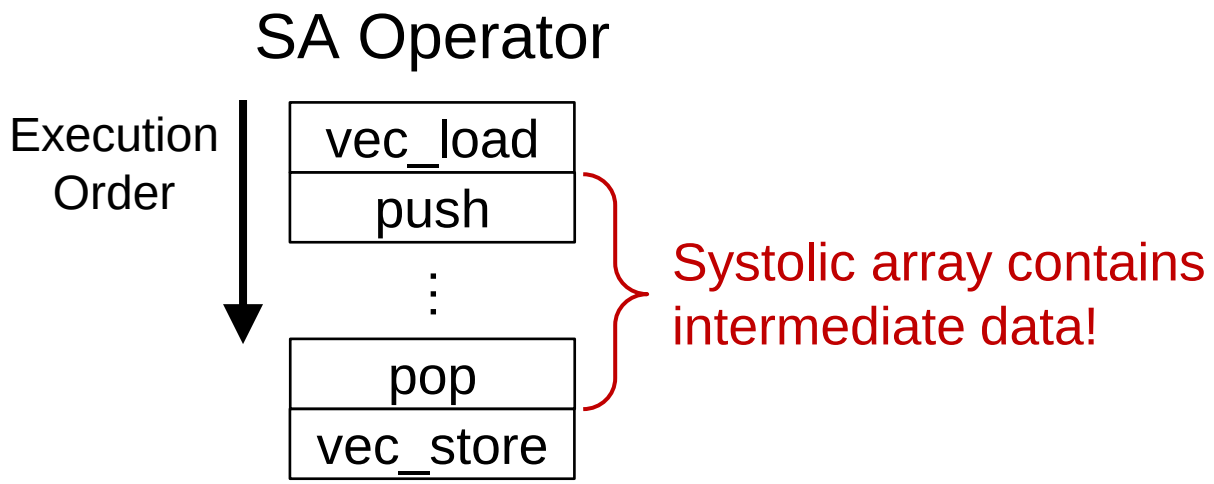


VU Preemption at Instruction Granularity

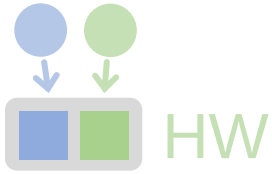


A Lightweight Context Switch Mechanism for Tensor Operator Preemption

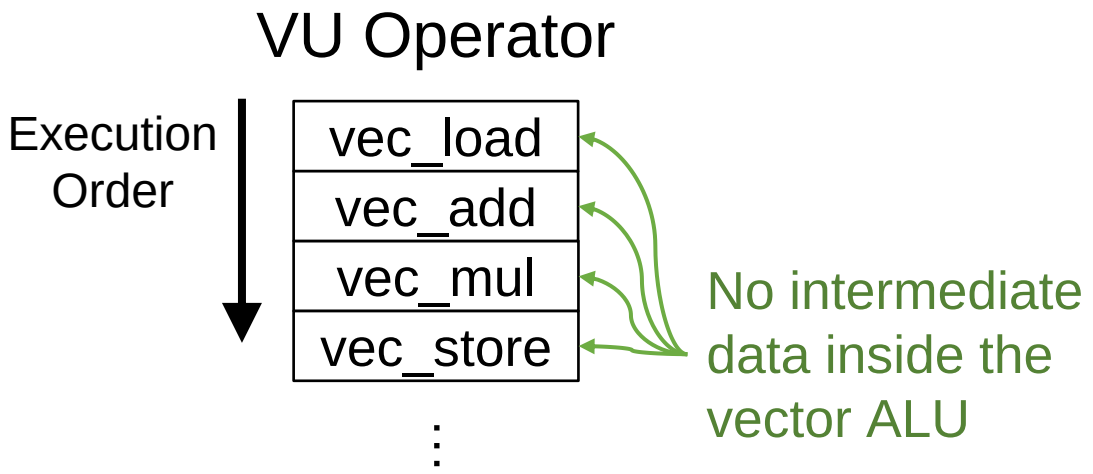


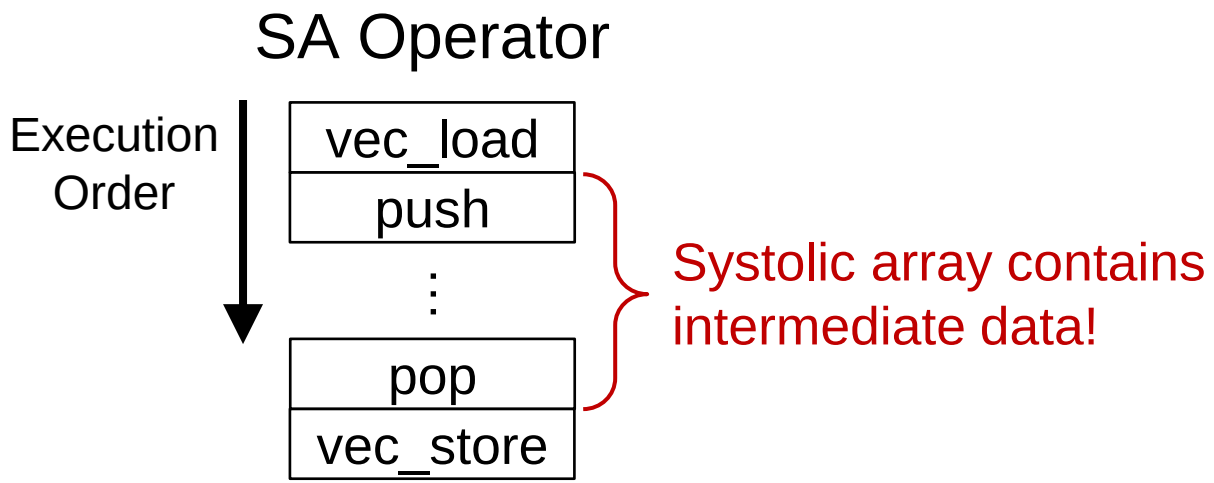


VU Preemption at Instruction Granularity



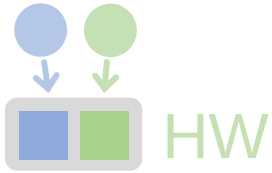
A Lightweight Context Switch Mechanism for Tensor Operator Preemption



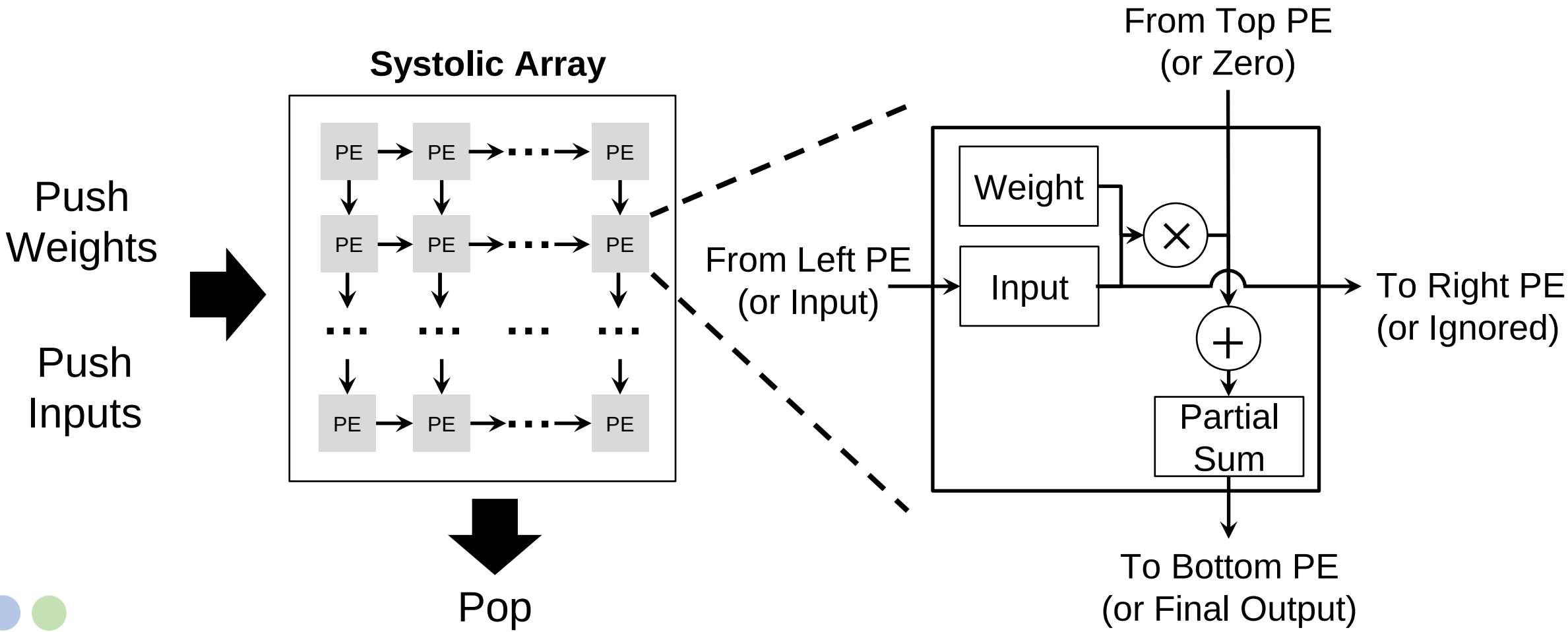


VU Preemption at Instruction Granularity

SA Preemption Requires Special Treatment

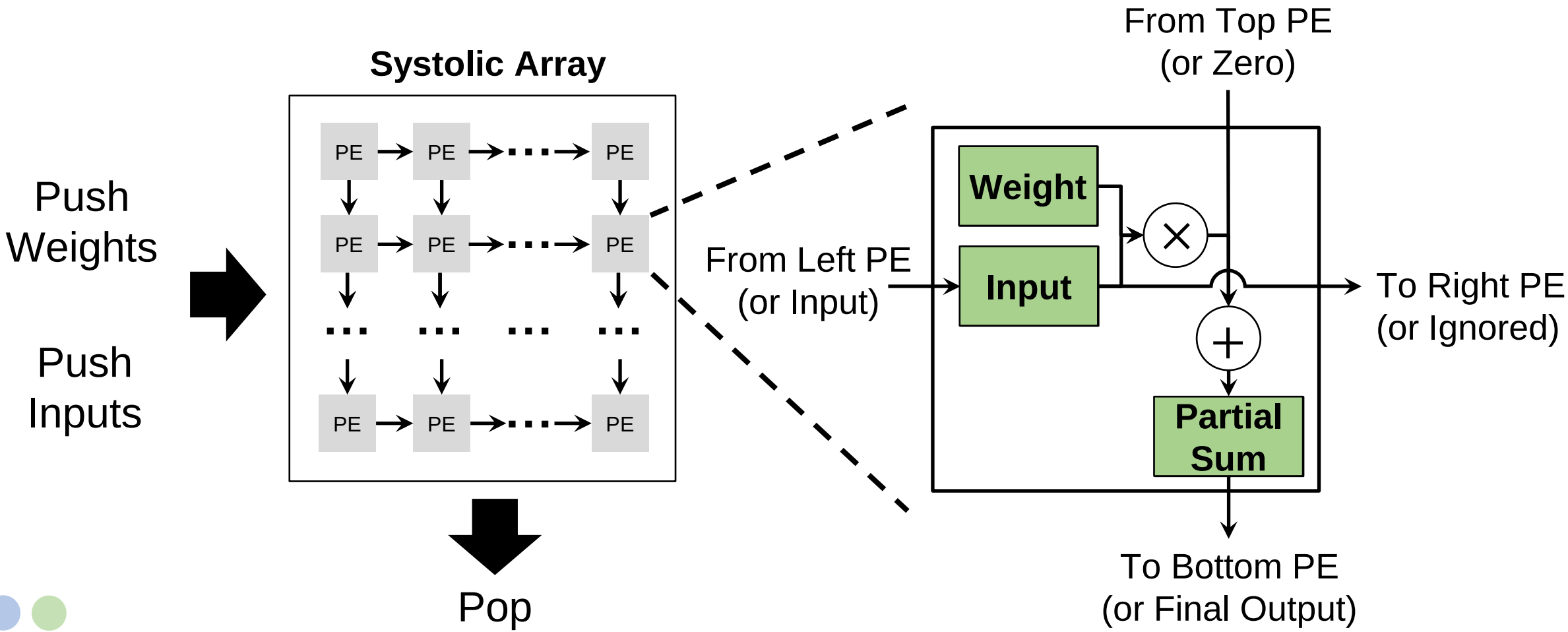


Lightweight Systolic Array Context Switching



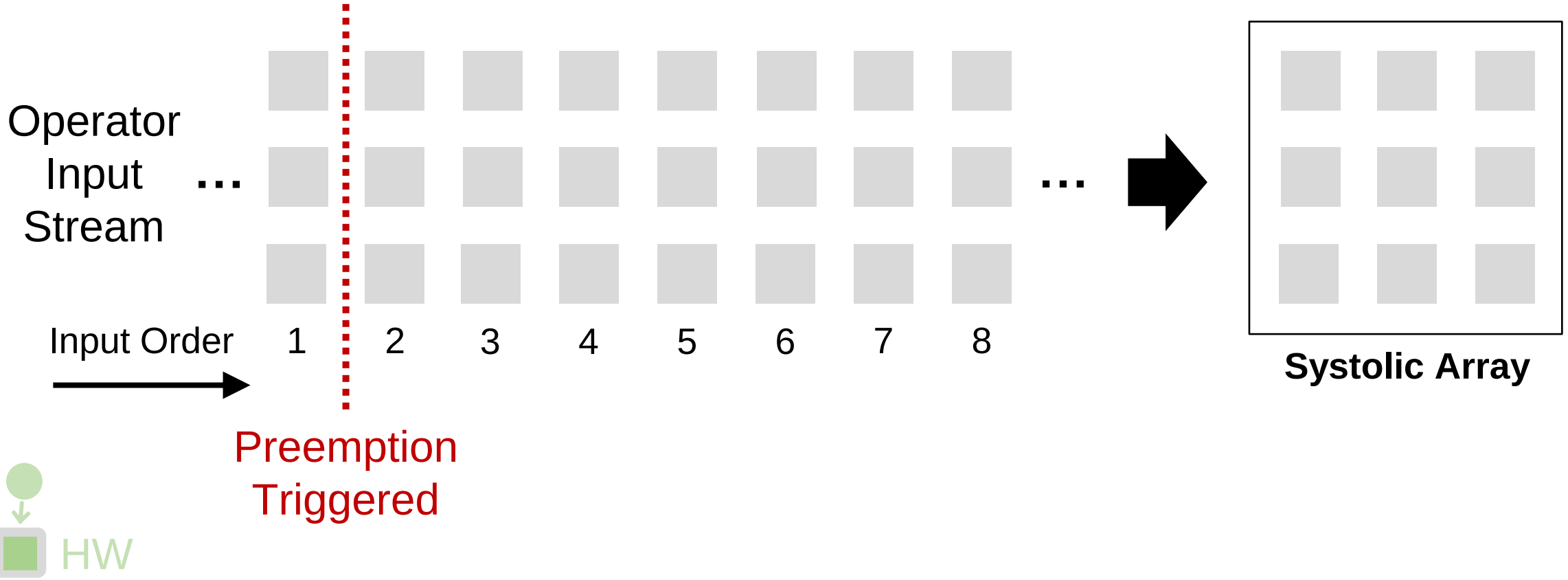
HW

Lightweight Systolic Array Context Switching

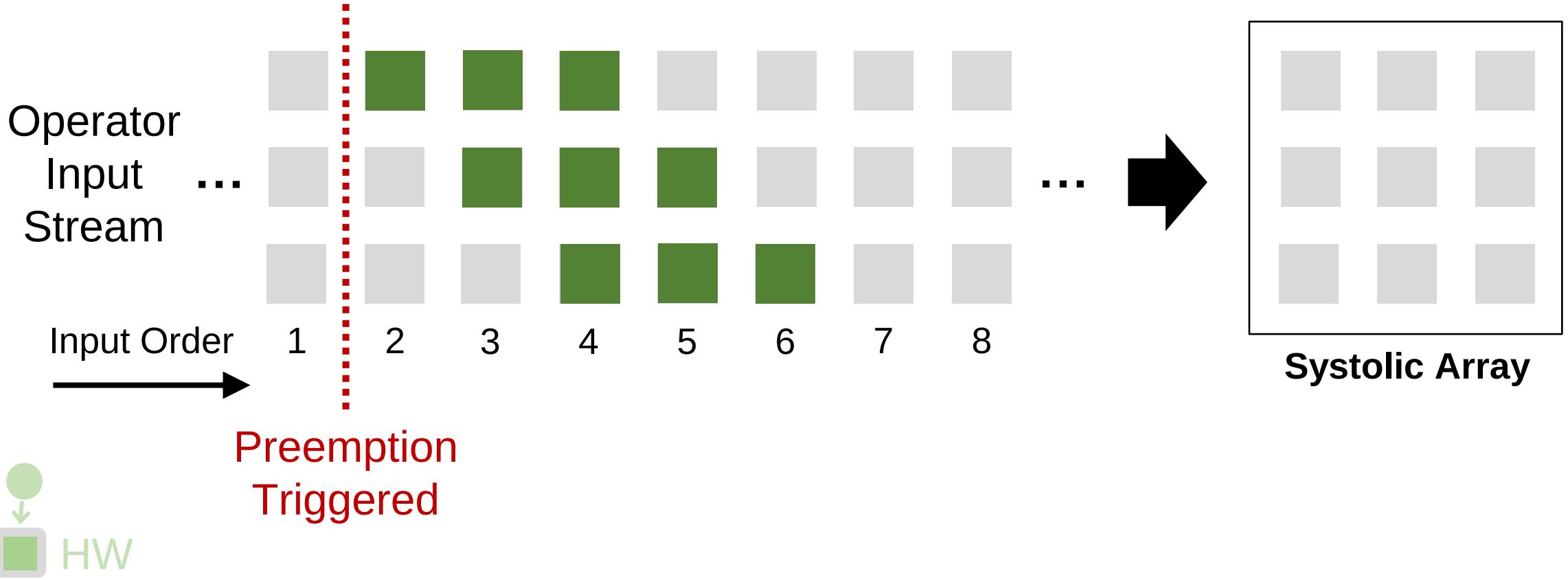


HW

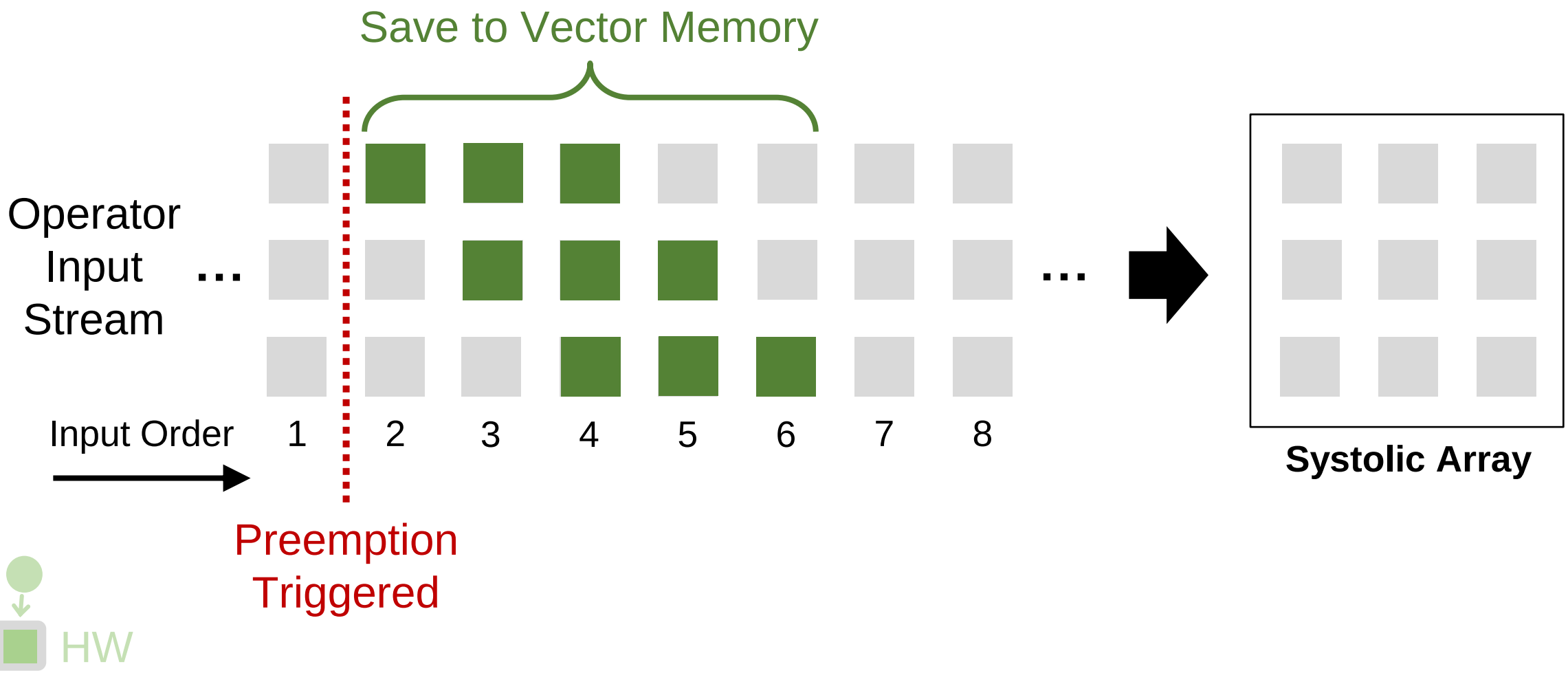
Lightweight Systolic Array Context Switching



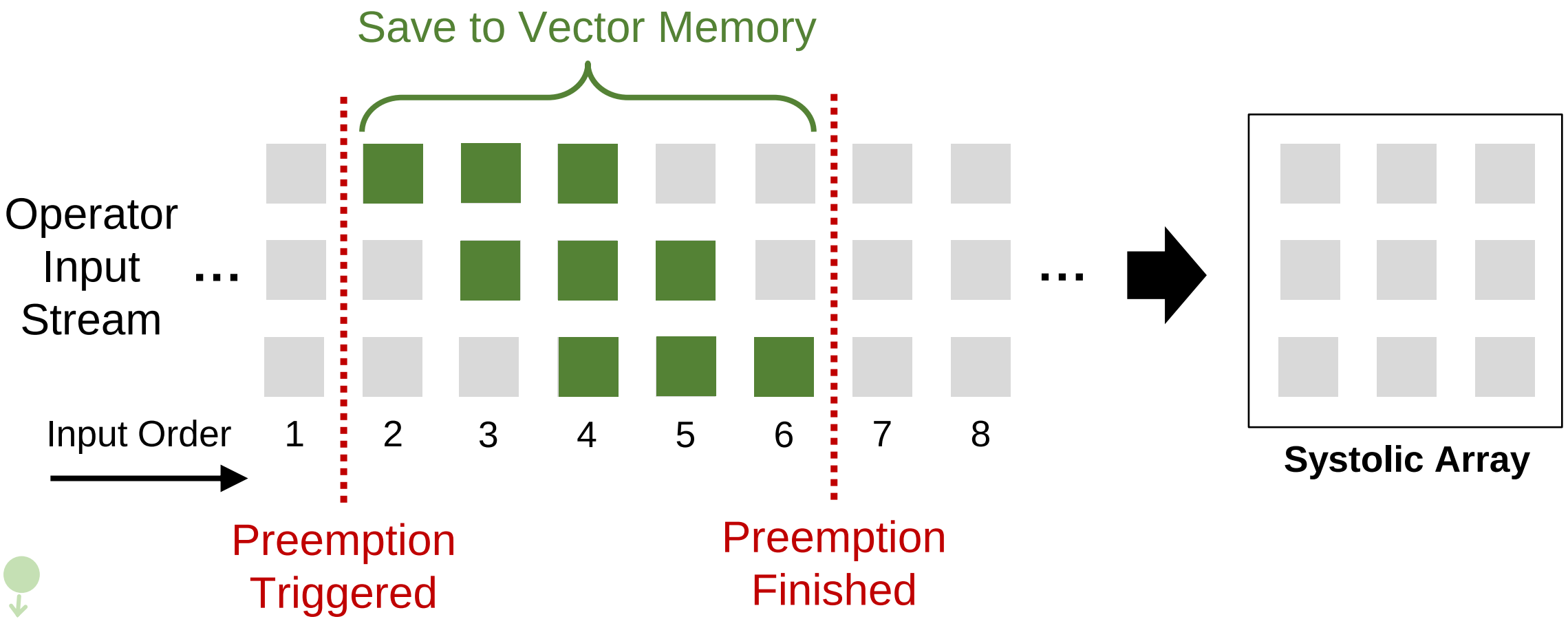
Lightweight Systolic Array Context Switching



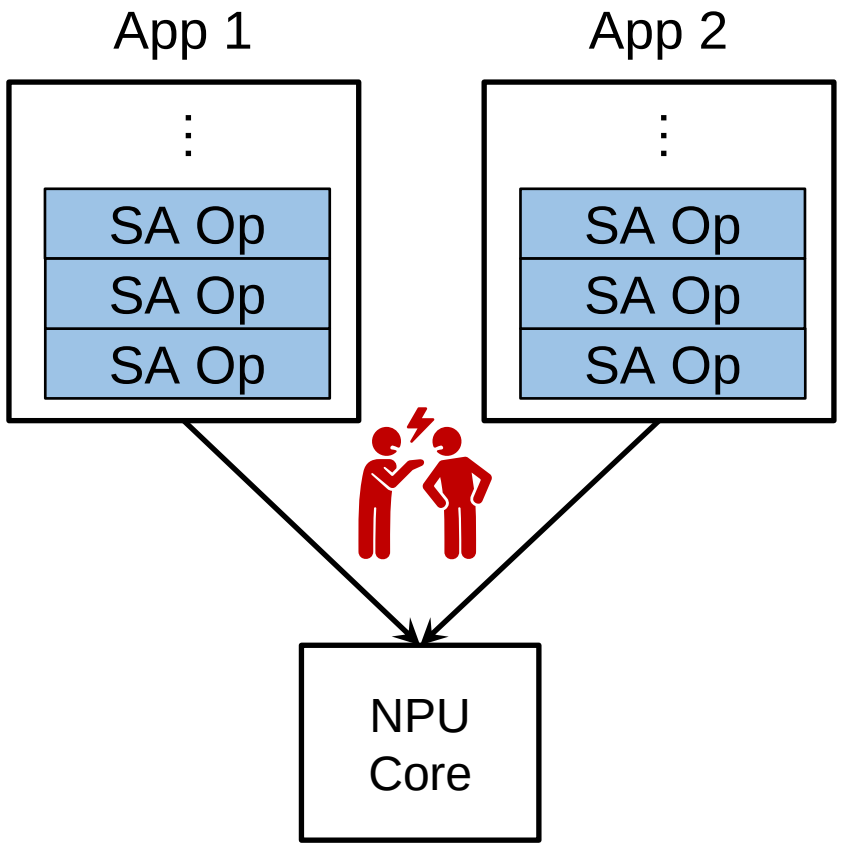
Lightweight Systolic Array Context Switching



Lightweight Systolic Array Context Switching

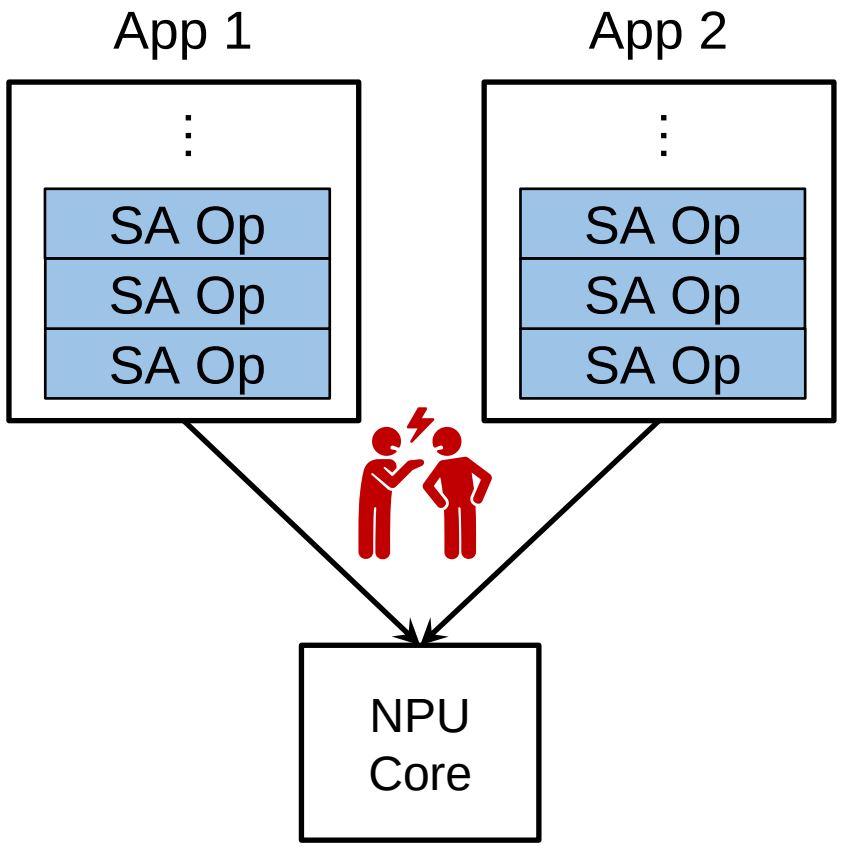


Random Collocation of DNN Workloads Incurs Significant Resource Contention



SW

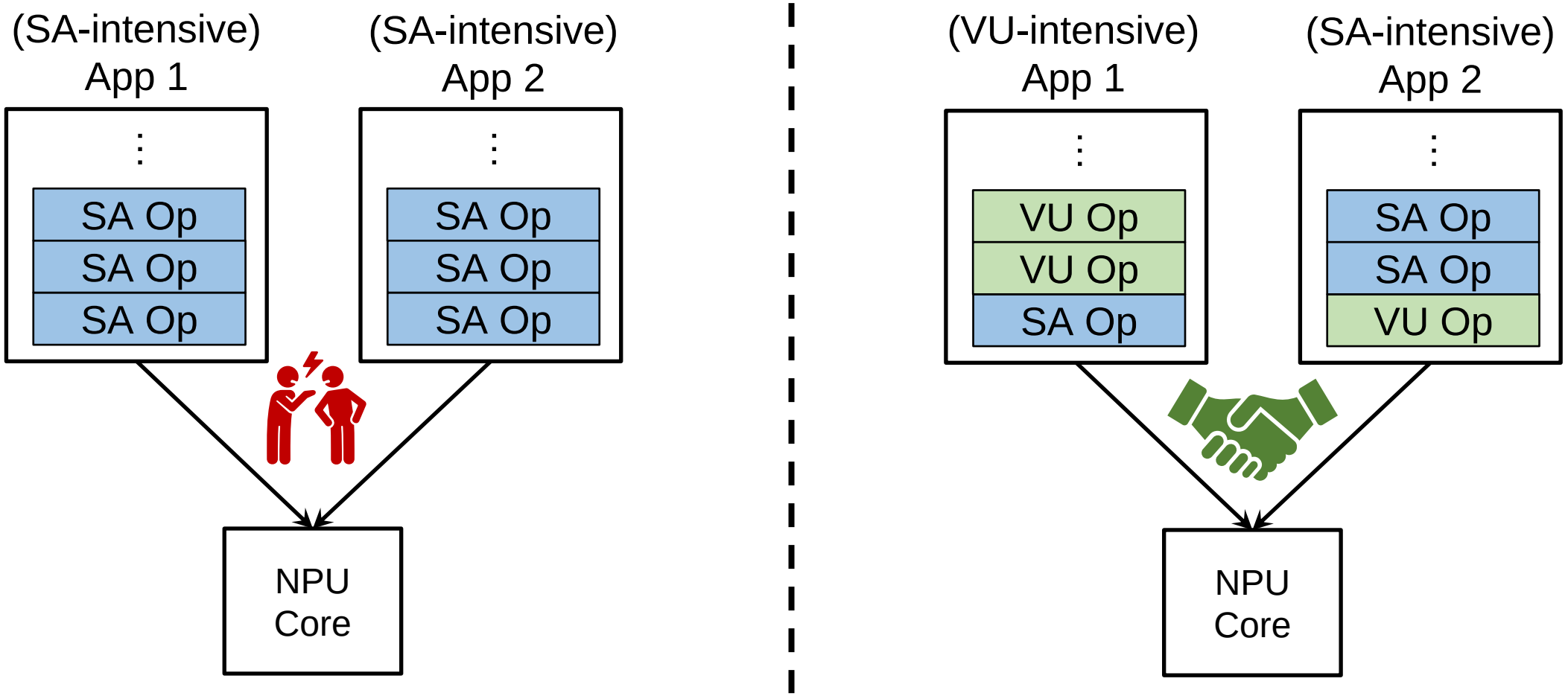
Random Collocation of DNN Workloads Incurs Significant Resource Contention



SW

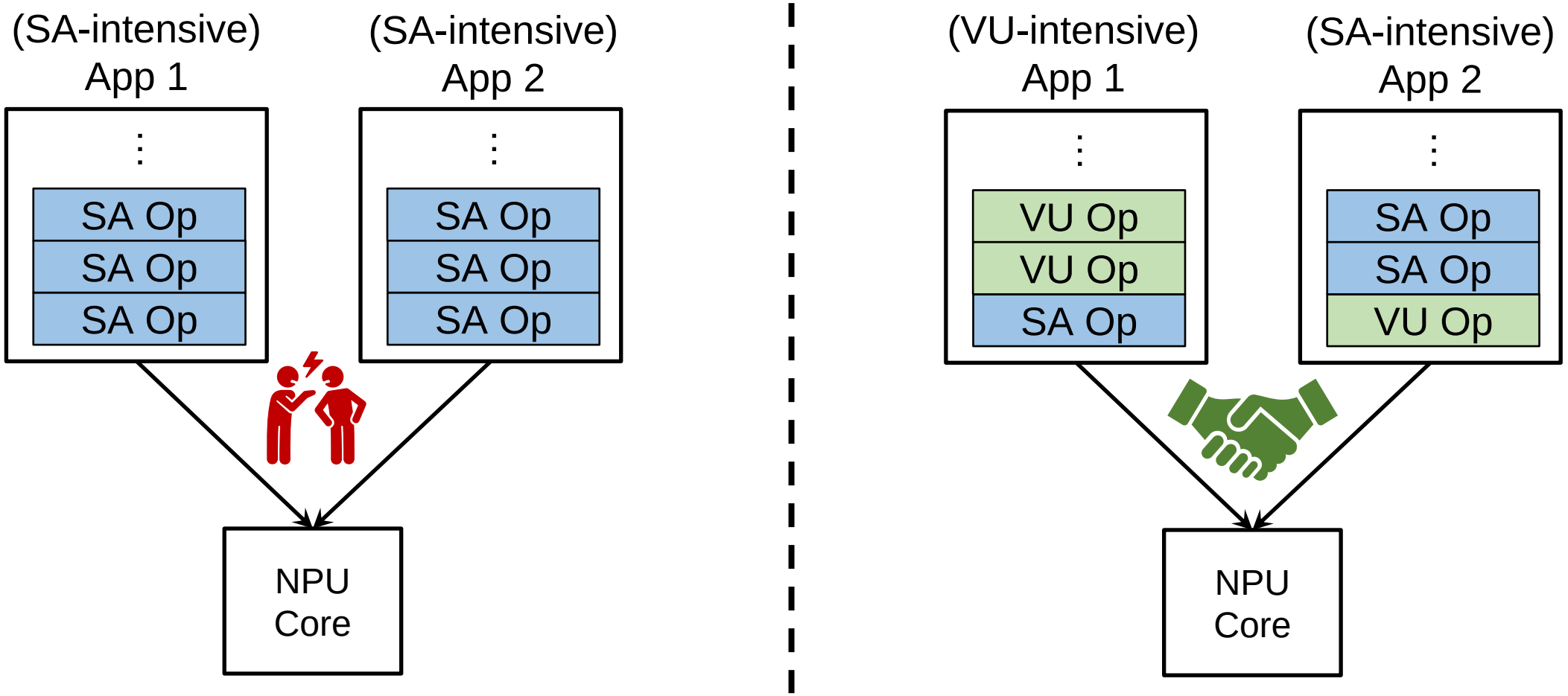
DNN Workloads May Demand for the Same Type of Resource

Random Collocation of DNN Workloads Incurs Significant Resource Contention



SW

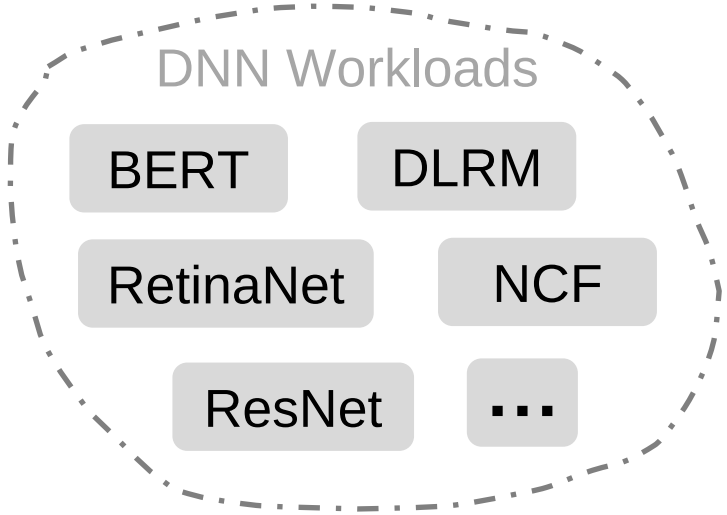
Random Collocation of DNN Workloads Incurs Significant Resource Contention



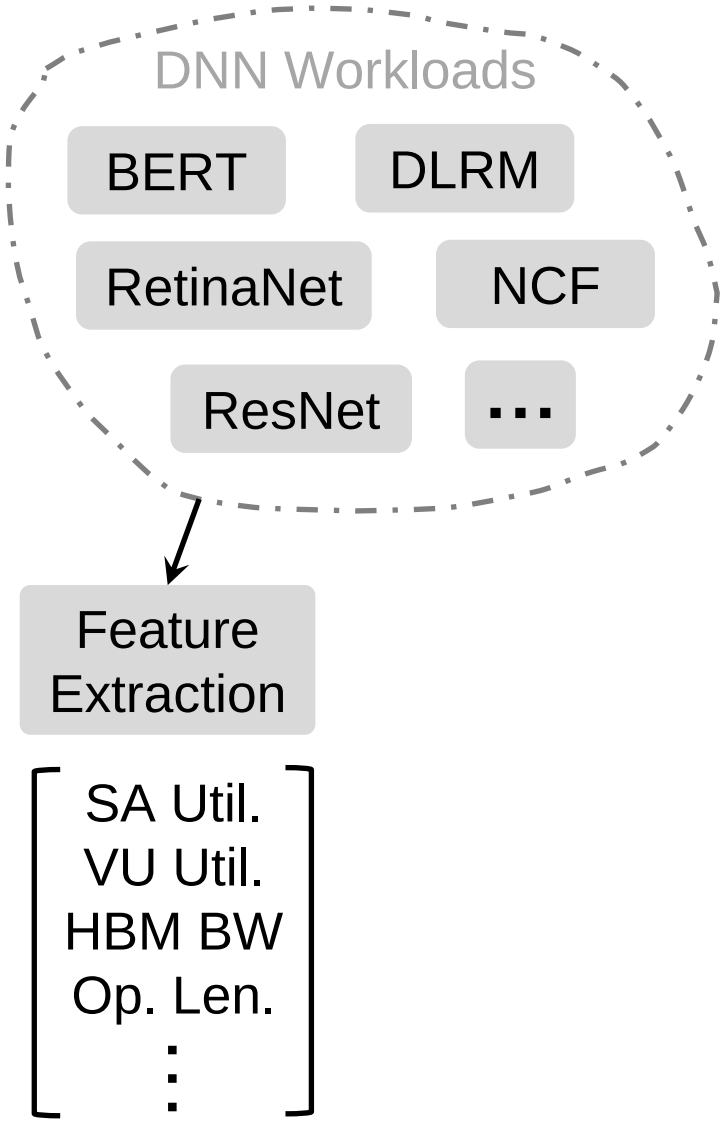
SW

Identifying Workloads Without Resource Contention Is Not Easy

Identifying Compatible Workloads with Clustering-based Collocation Mechanism

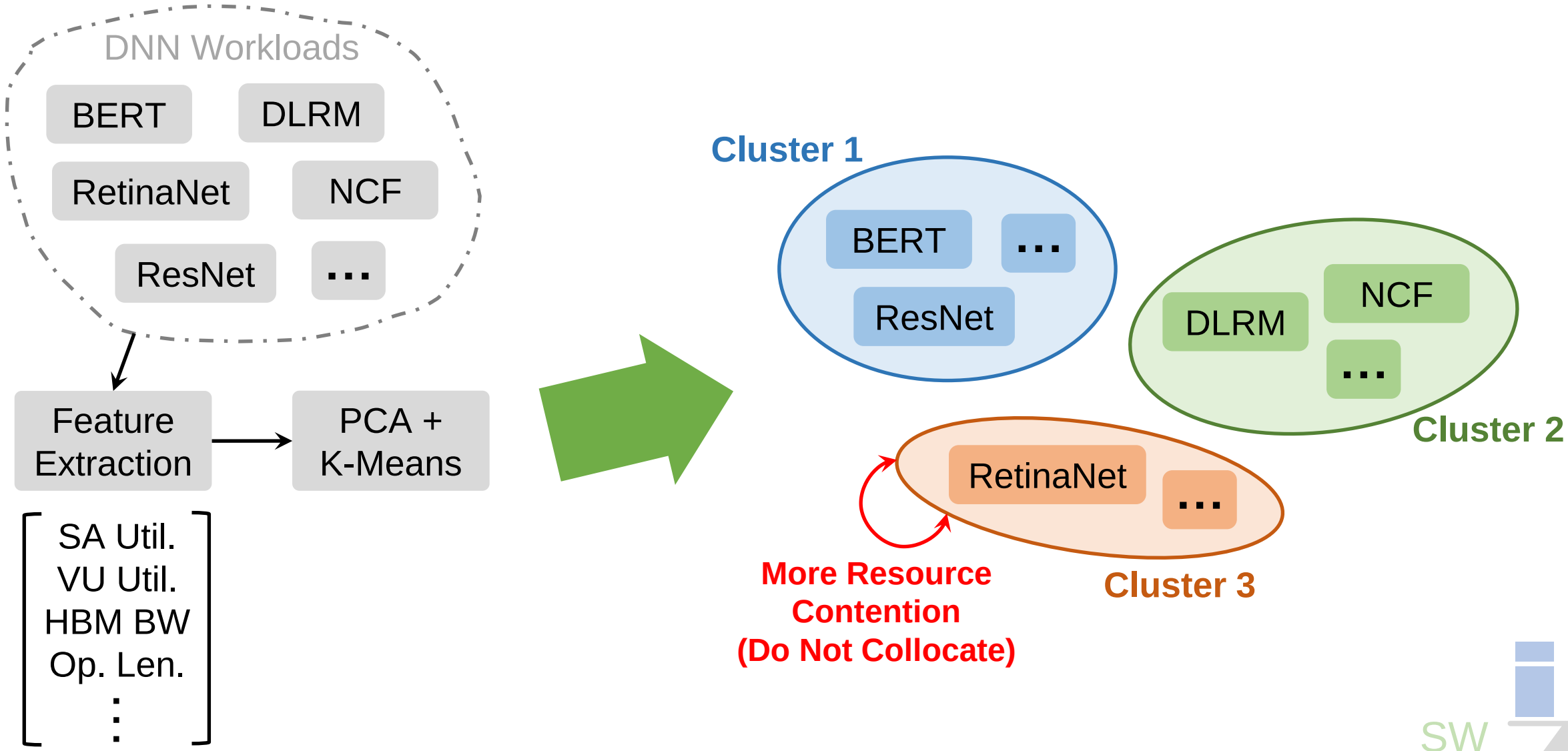


Identifying Compatible Workloads with Clustering-based Collocation Mechanism

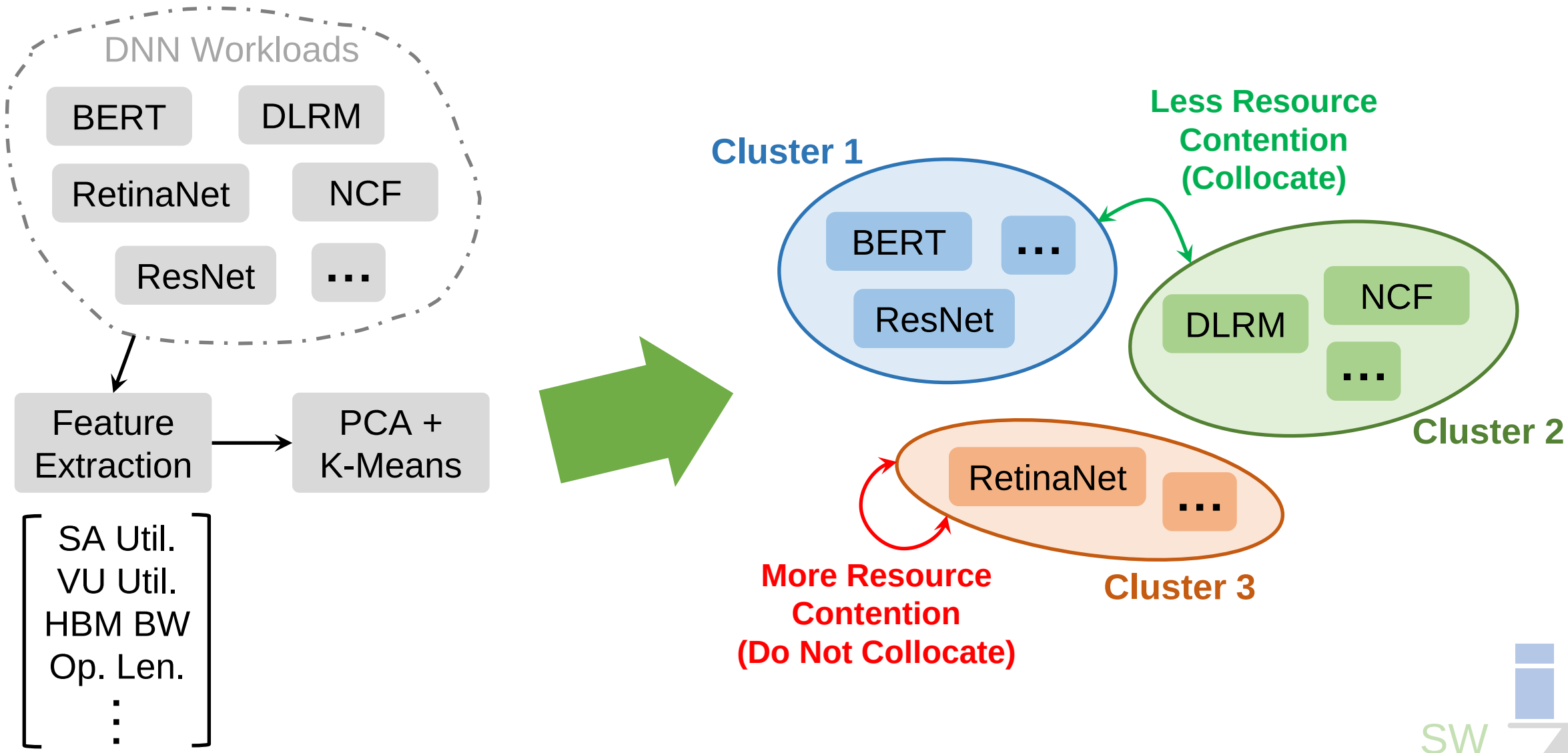


SW

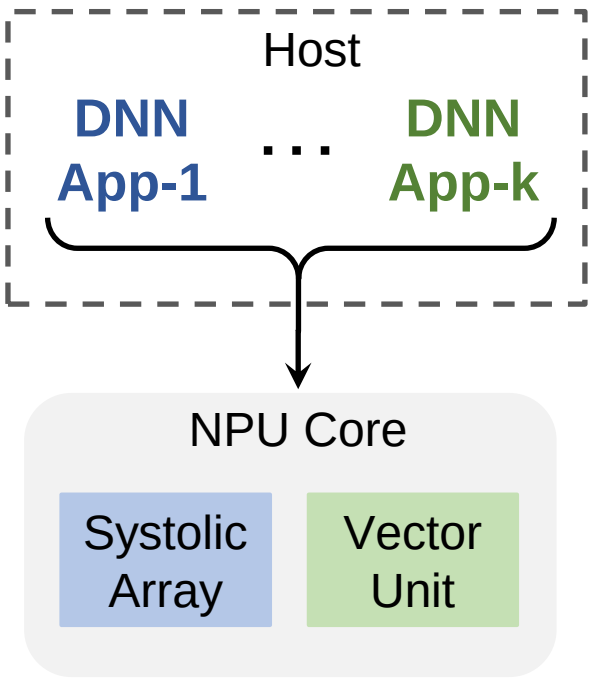
Identifying Compatible Workloads with Clustering-based Collocation Mechanism



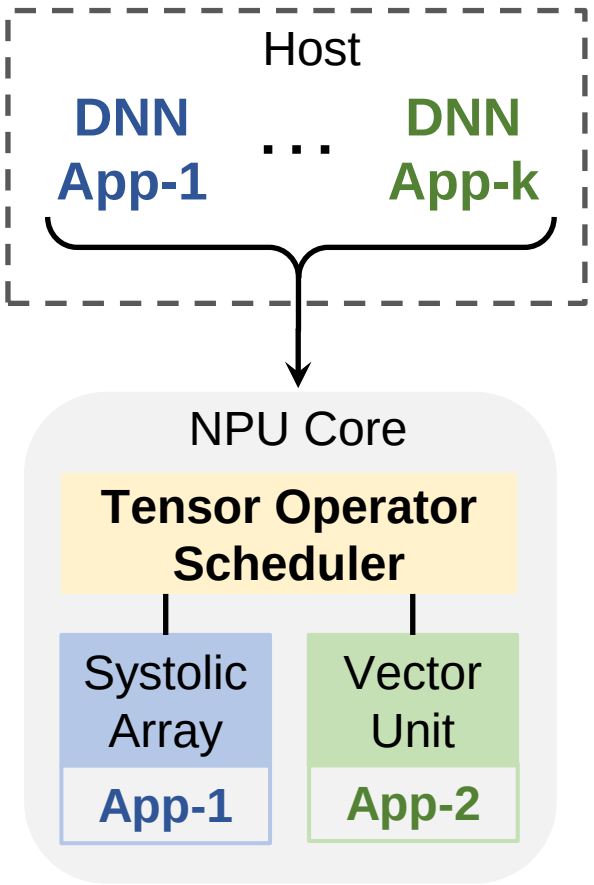
Identifying Compatible Workloads with Clustering-based Collocation Mechanism

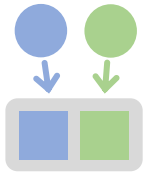


Put It All Together



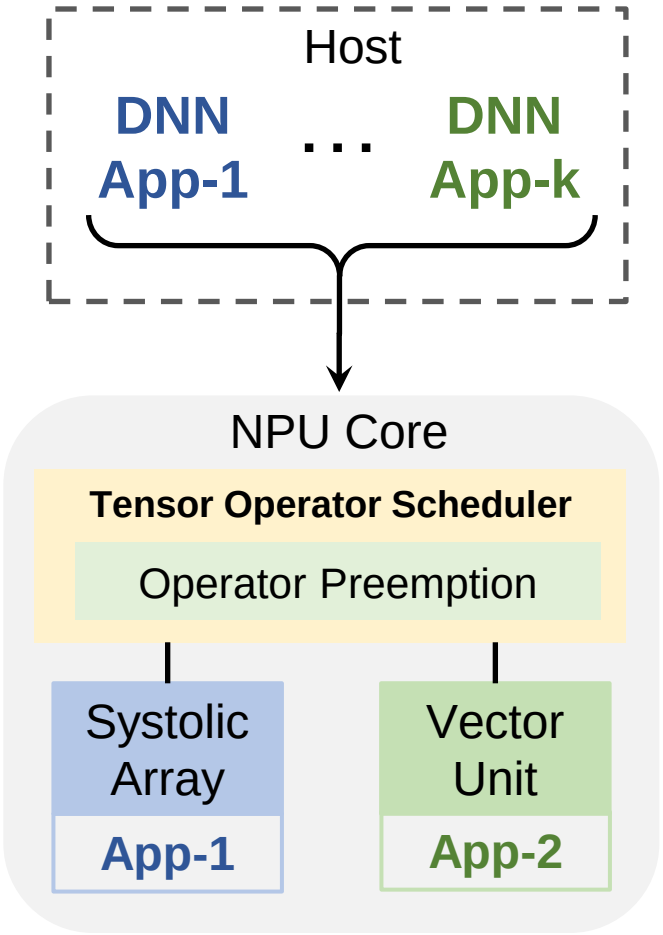
Put It All Together

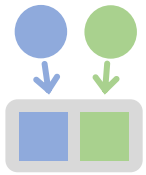




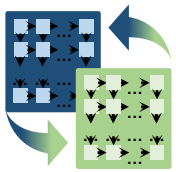
Architectural Support for SA/VU-level Resource Sharing

Put It All Together



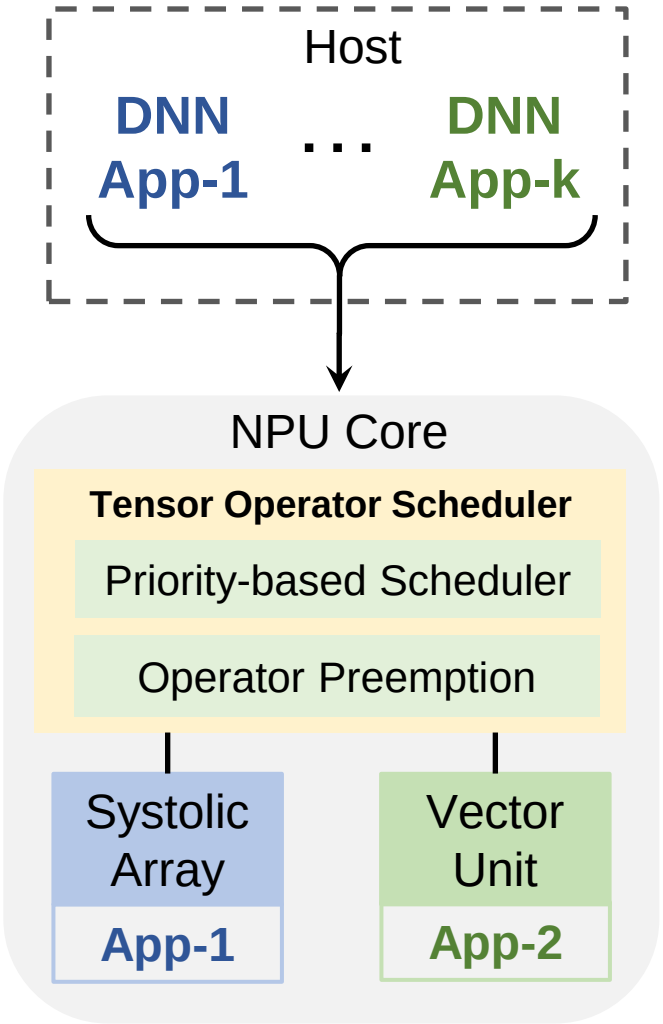


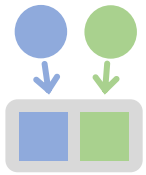
Architectural Support for SA/VU-level Resource Sharing



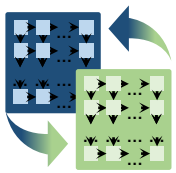
Lightweight Tensor Operator Preemption

Put It All Together

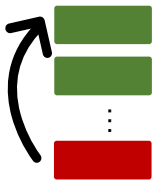




Architectural Support for SA/VU-level Resource Sharing

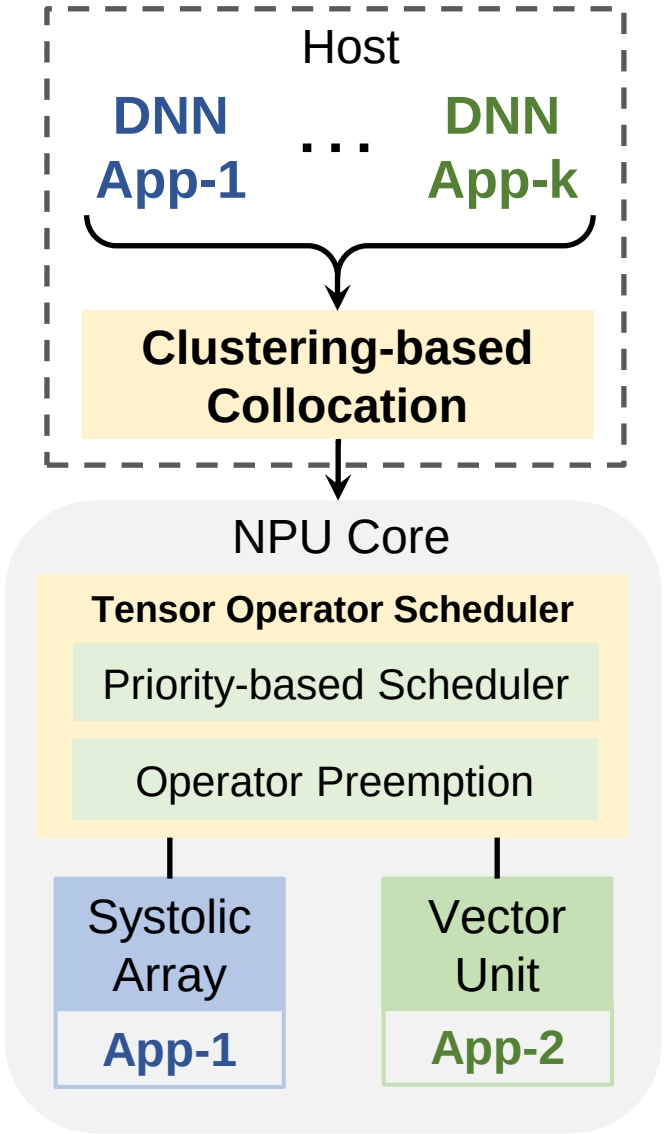


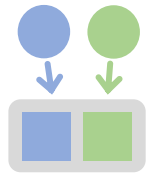
Lightweight Tensor Operator Preemption



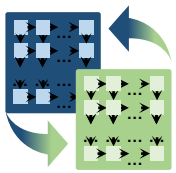
Priority-based Scheduler

Put It All Together

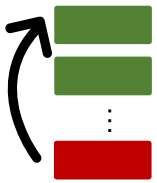




Architectural Support for SA/VU-level Resource Sharing



Lightweight Tensor Operator Preemption



Priority-based Scheduler

Clustering-based Collocation Mechanism

Evaluation

Implementation

Trace-driven simulator
based on Google TPU

Evaluation

Implementation — Trace-driven simulator based on Google TPU

Benchmarks — MLPerf v2.1 and TPU Reference Models

Evaluation

Implementation

Trace-driven simulator
based on Google TPU

Benchmarks

MLPerf v2.1 and
TPU Reference Models

Experimental Setup

- **PMT:** Preemptive Multi-tasking at NPU core-level

Evaluation

Implementation

Trace-driven simulator
based on Google TPU

Benchmarks

MLPerf v2.1 and
TPU Reference Models

Experimental Setup

- **PMT:** Preemptive Multi-tasking at NPU core-level
- **V10-Base:** SA/VU-level scheduling w/o operator preemption

Evaluation

Implementation

Trace-driven simulator
based on Google TPU

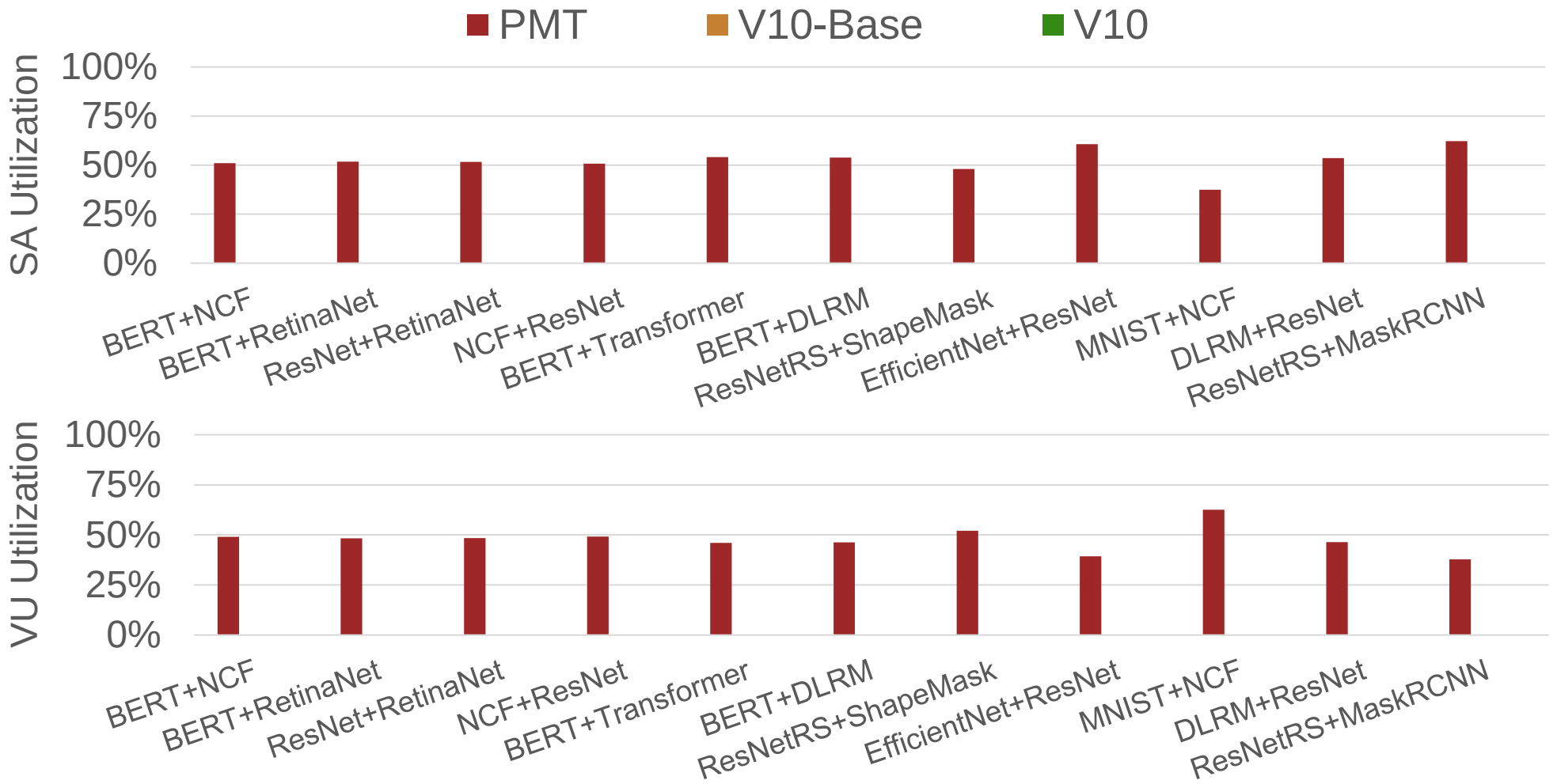
Benchmarks

MLPerf v2.1 and
TPU Reference Models

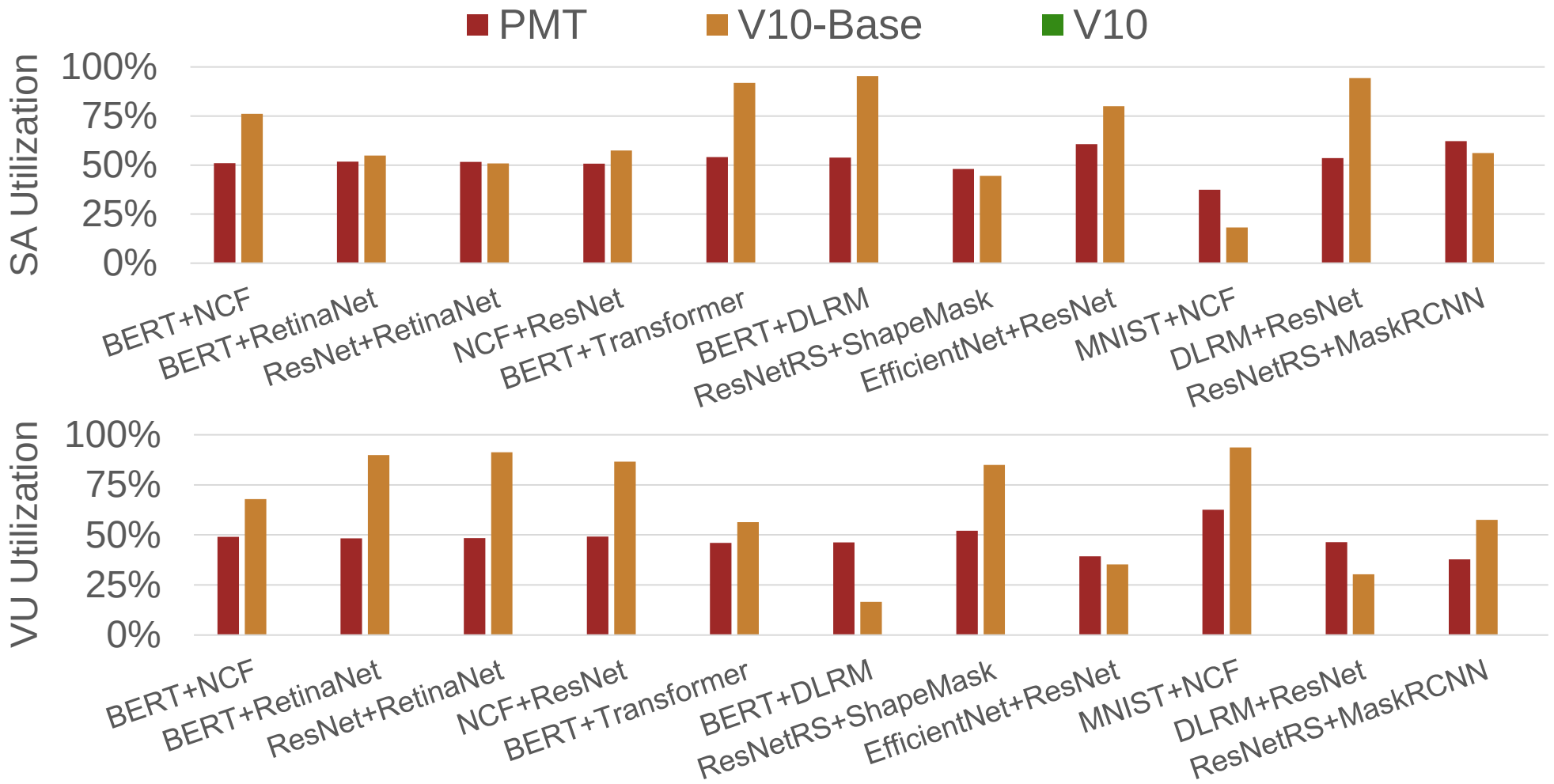
Experimental Setup

- **PMT:** Preemptive Multi-tasking at NPU core-level
- **V10-Base:** SA/VU-level scheduling w/o operator preemption
- **V10:** SA/VU-level scheduling w/ operator preemption

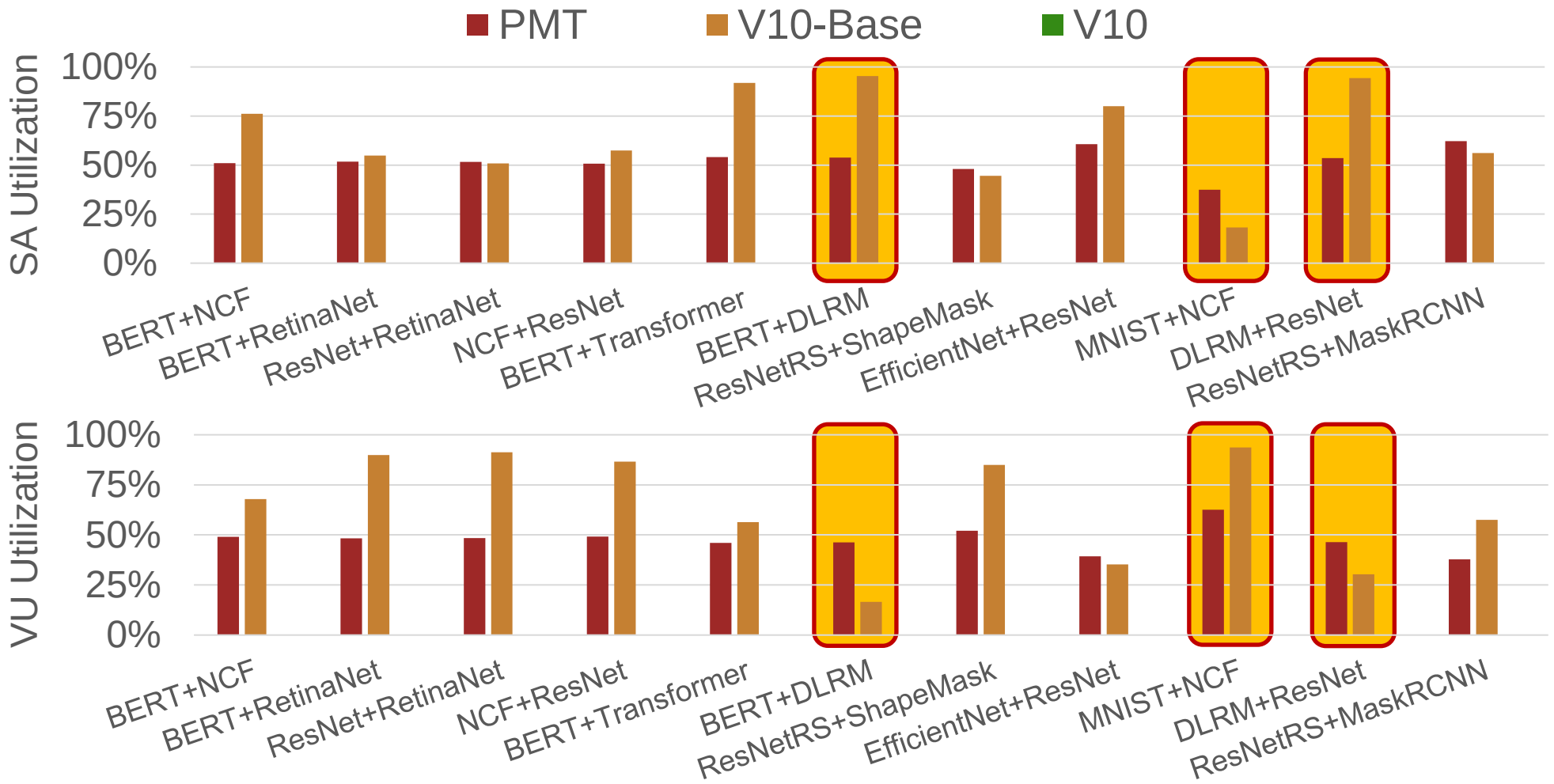
Utilization Improvement of NPU Cores with V10



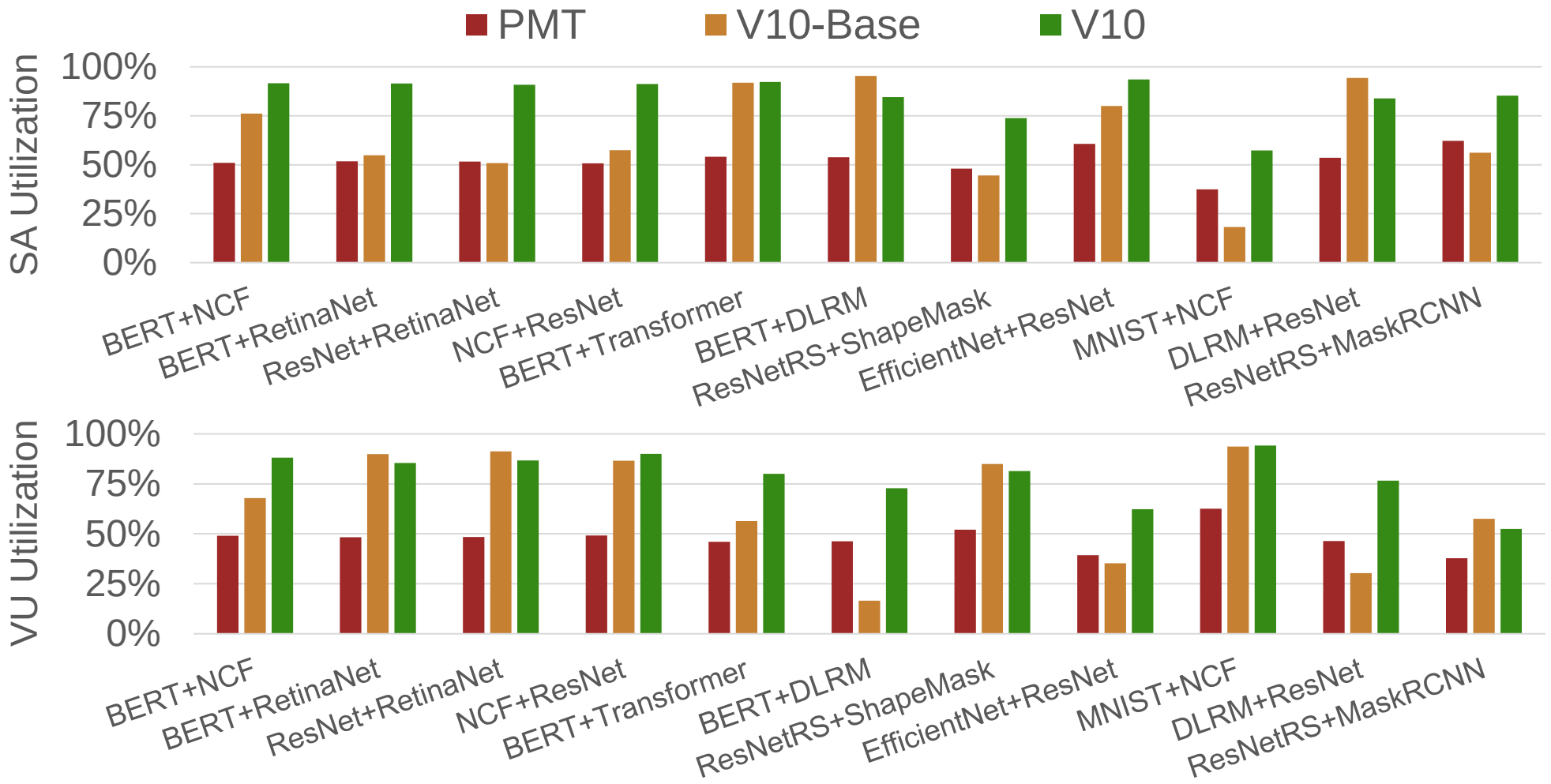
Utilization Improvement of NPU Cores with V10



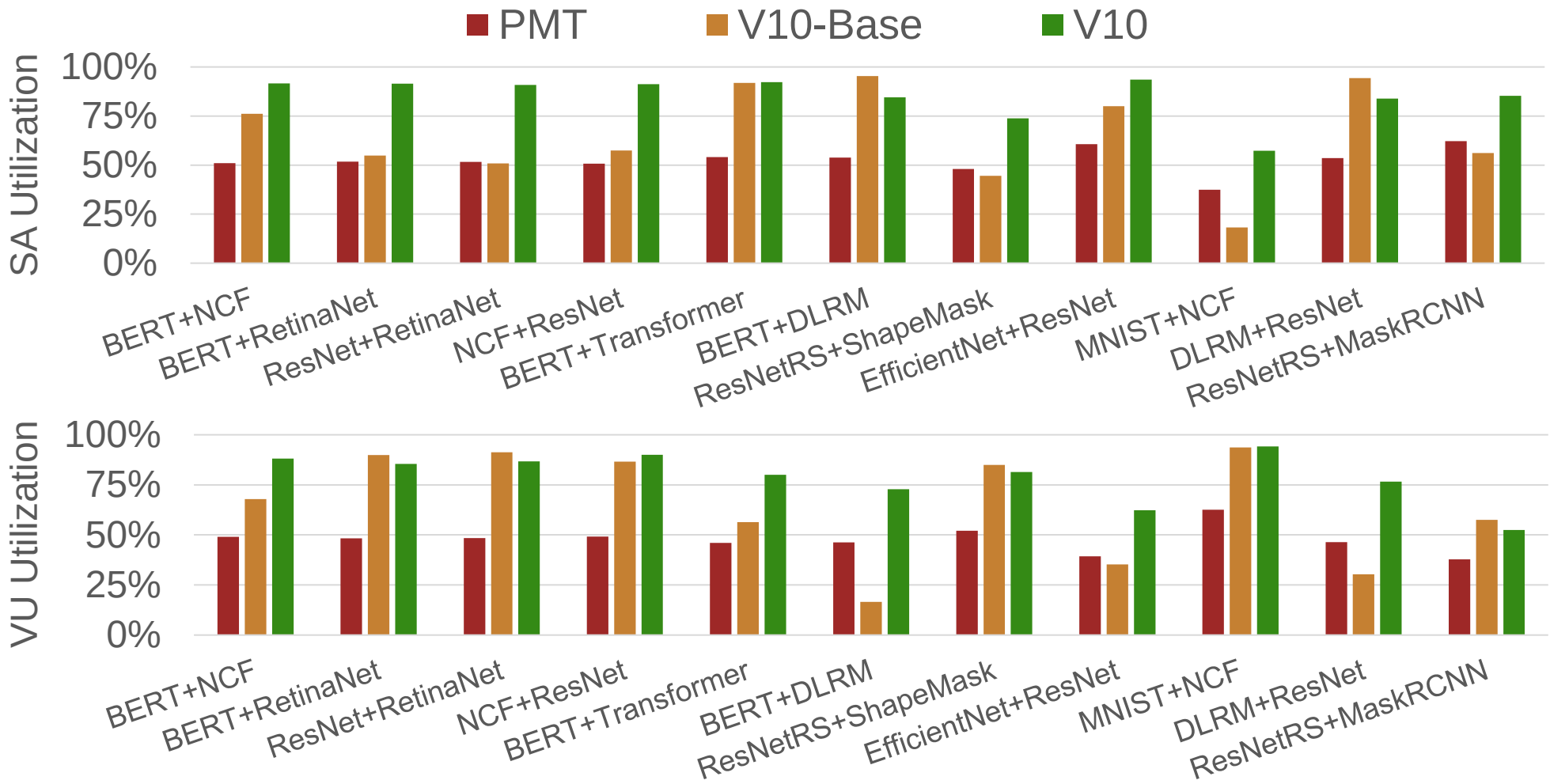
Utilization Improvement of NPU Cores with V10



Utilization Improvement of NPU Cores with V10

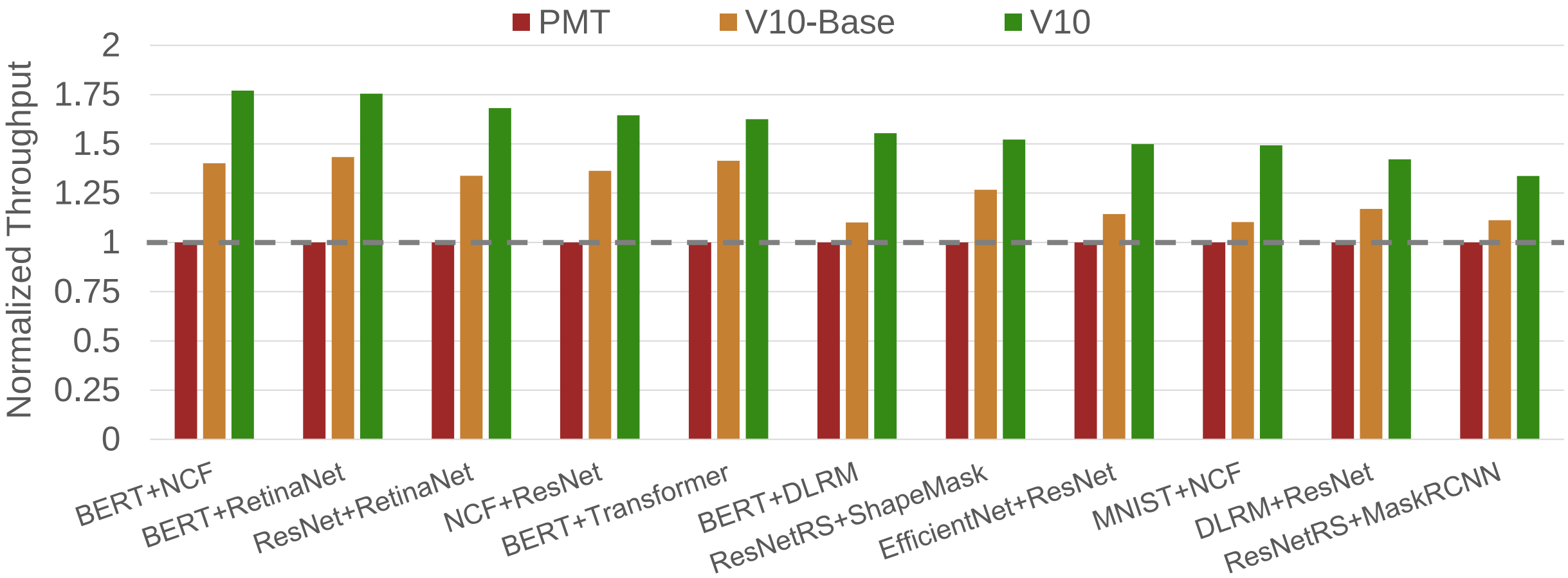


Utilization Improvement of NPU Cores with V10



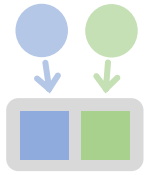
V10 Delivers 1.64x Utilization Improvement for NPU Cores

Throughput Improvement for DNN Inference Workloads with V10



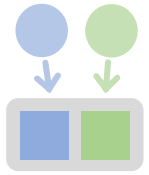
V10 Achieves 1.57x Throughput Improvement for Multi-tenant DNN Workloads

V10 Summary



Architectural Support for
Fine-grained NPU Sharing

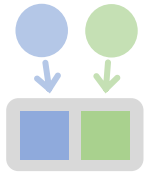
V10 Summary



Architectural Support for Fine-grained NPU Sharing

Clustering-based Collocation Mechanism for ML Workloads

V10 Summary



Architectural Support for Fine-grained NPU Sharing

Clustering-based Collocation Mechanism for ML Workloads

Improved NPU Utilization by 1.64x

Thank you!

Yuqi Xue
yuqixue2@illinois.edu

Yiqi Liu Lifeng Nai Jian Huang

Systems Platform Research Group

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN