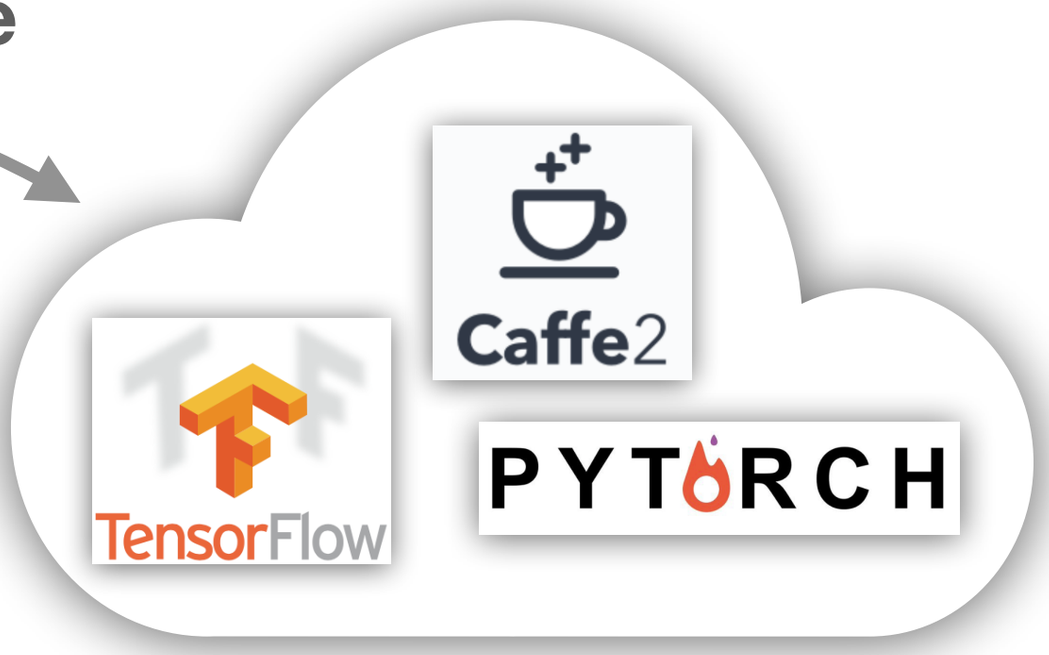
Couper: DNN Model Slicing for Video Analytics Containers at the Edge

Ke-Jou (Carol) Hsu
Ketan Bhardwaj
Ada Gavrilovska

Video analytics applications are in high demand

Raw image frame

Analysis result

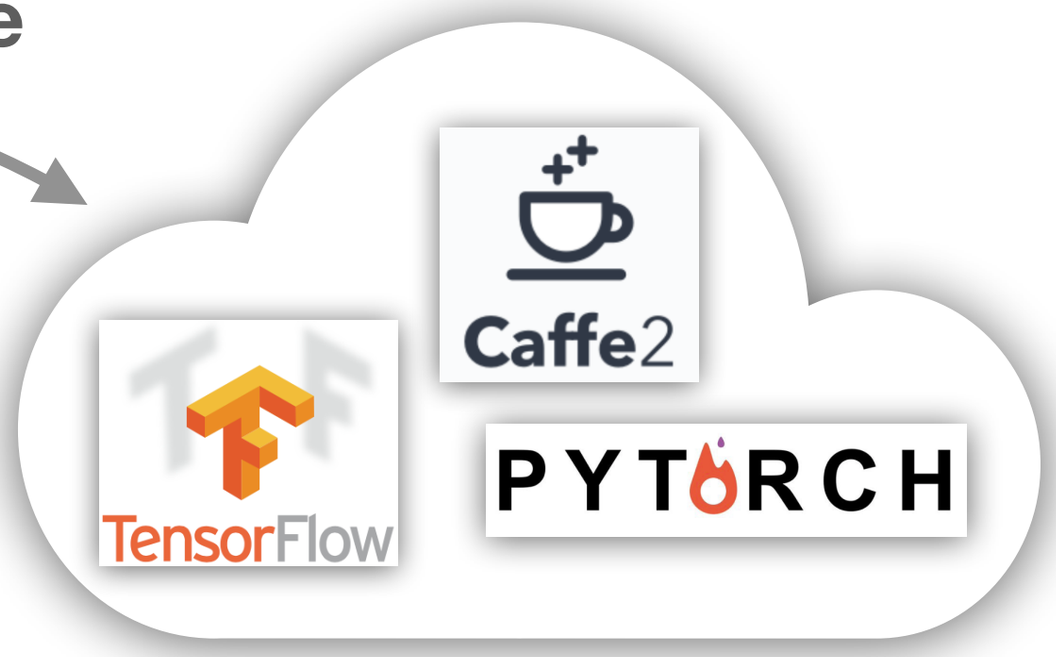


Video analytics applications are in high demand

Raw image frame

Communication cost

Analysis result

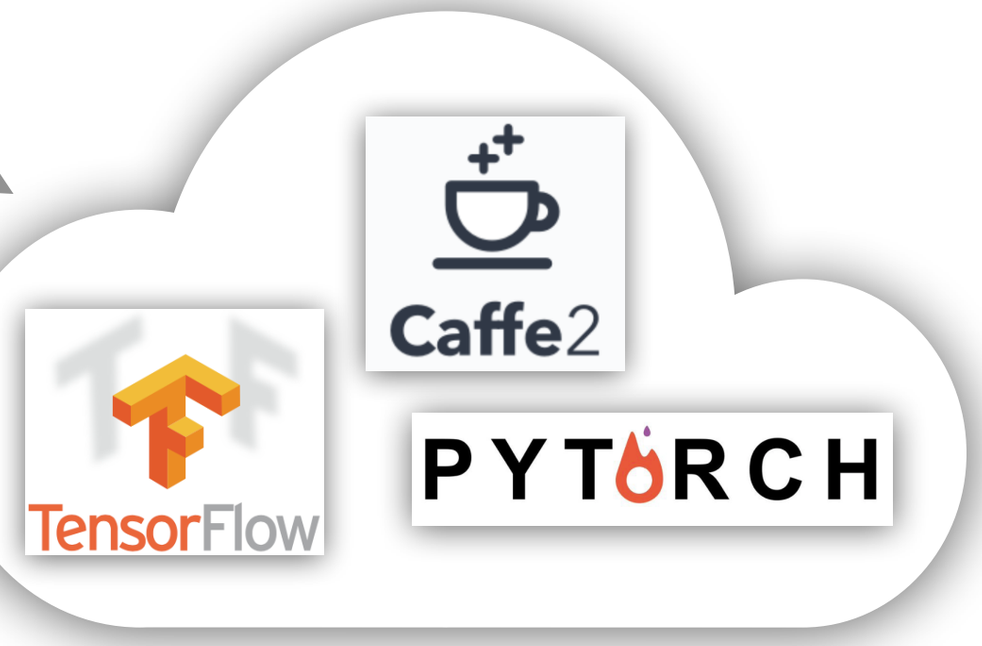


Video analytics applications are in high demand

Raw image frame

Communication cost

Analysis result



Video analytics application may face great **performance degradation** because of its **data-intensive** and **latency-sensitive** workload

Edge's proximity benefit can help!

Edge computing brings benefits:

- **Higher computing resource than client**
- **Reduce communication cost, lower processing latencies, higher processing rates, …**
- **Flexible service deployment**

How does video analytics application work with edge?

How does video analytics application work with edge?

Deep neural network (DNN)

How does video analytics application work with edge?

Deep neural network (DNN)

↓

High accuracy and famous

How does video analytics application work with edge?

Deep neural network (DNN)

High accuracy and famous

Computation-intensive workload

How does video analytics application work with edge?

Deep neural network (DNN)

High accuracy and famous

Computation-intensive workload

Model	VGG 16	MobileNet V2 1.4	ResNet V2 50	Inception V3	Inception ResNet V2	NASNet 331	PNASNet 331
Released Time	**2014 Sep**	2018 Jan	2016 Jul	2016 Jul	2016 Aug	2018 Apr	**2018 Jul**
Top-1 Accuracy	**71.5**	74.9	75.6	78.0	80.4	82.7	**82.9**
# Operators	**54**	155	205	788	871	1265	**939**

Accuracy increases, so does model complexity

How does video analytics application work with edge?

Deep neural network (DNN)

- Google, Cliff Young
  (Linley processor conference 2018)

How does video analytics application work with edge?

Deep neural network (DNN)

- Google, Cliff Young
  (Linley processor conference 2018)

Single type of device cannot fit **every DNN**,
more accurate DNNs require more resource

How does video analytics application work with edge?

Deep neural network (DNN)

Client -> Edge -> Cloud

How does video analytics application work with edge?

Deep neural network (DNN)

Client -> Edge -> Cloud

Diverse specification and network distance

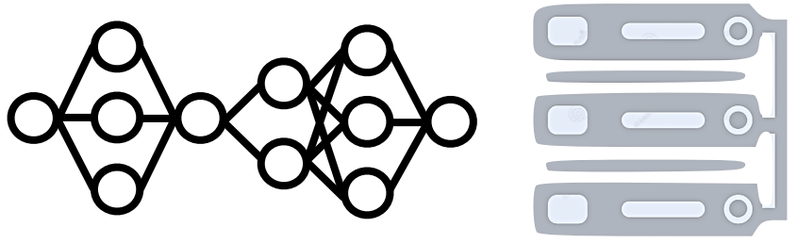
Bringing out edge's benefit is not easy

Bringing out edge's benefit is not easy

If edge cannot run whole DNN:

Bringing out edge's benefit is not easy

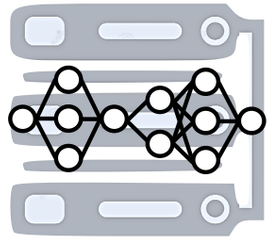
If edge cannot run whole DNN:



Optimize DNN for edge

Bringing out edge's benefit is not easy

If edge cannot run whole DNN:

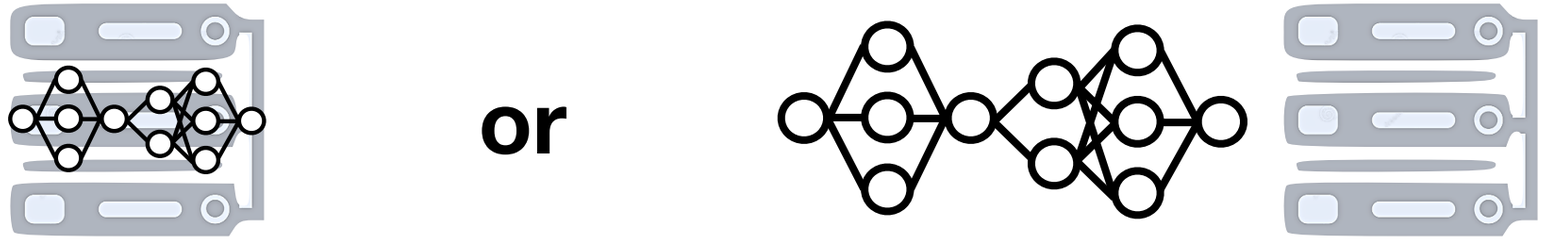


Optimize DNN for edge

Bringing out edge's benefit is not easy

DNN ? Client -> Edge -> Cloud

If edge cannot run whole DNN:



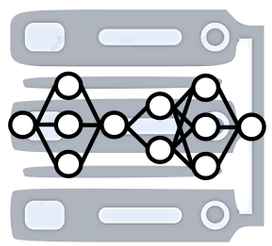
Optimize DNN for edge

or

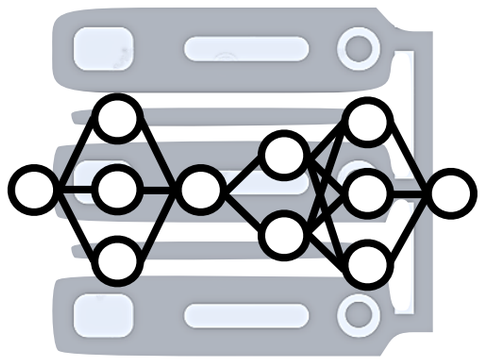
Bring specific edge for DNN

Bringing out edge's benefit is not easy

If edge cannot run whole DNN:



or



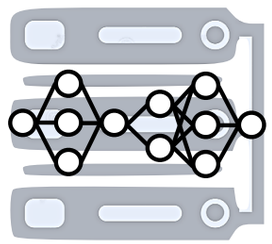
Optimize DNN for edge

Bring specific edge for DNN

Bringing out edge's benefit is not easy

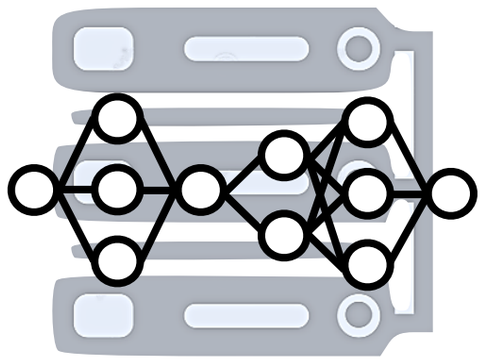
DNN ? Client -> Edge -> Cloud

If edge cannot run whole DNN:



Optimize DNN for edge

or



Bring specific edge for DNN

These two methods are relatively **time- and money-consuming**
and turns to be **impractical** for
rapid growth of DNNs and **diverse** and **shared** edge environment

Problem Statement

This is a multi-dimensional problem:

1. **Heterogeneous computing resource** on client-edge-cloud.
2. **Various** compute-intensive **DNN** models
3. **No single deployment** meets users' expectation **forever**

Problem Statement

This is a multi-dimensional problem:

1. **Heterogeneous computing resource** on client-edge-cloud.
2. **Various** compute-intensive **DNN** models
3. **No single deployment** meets users' expectation **forever**

Given a DNN and an edge,

How can we deploy the model with good performance?

Problem Statement

This is a multi-dimensional problem:

1. **Heterogeneous computing resource** on client-edge-cloud.
2. **Various** compute-intensive **DNN** models
3. **No single deployment** meets users' expectation **forever**

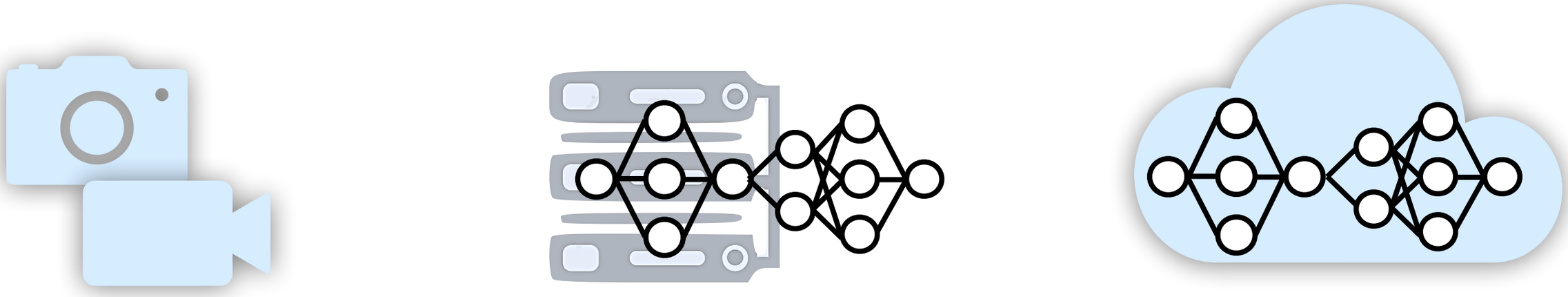
Given a DNN and an edge,

How can we deploy the model with good performance?

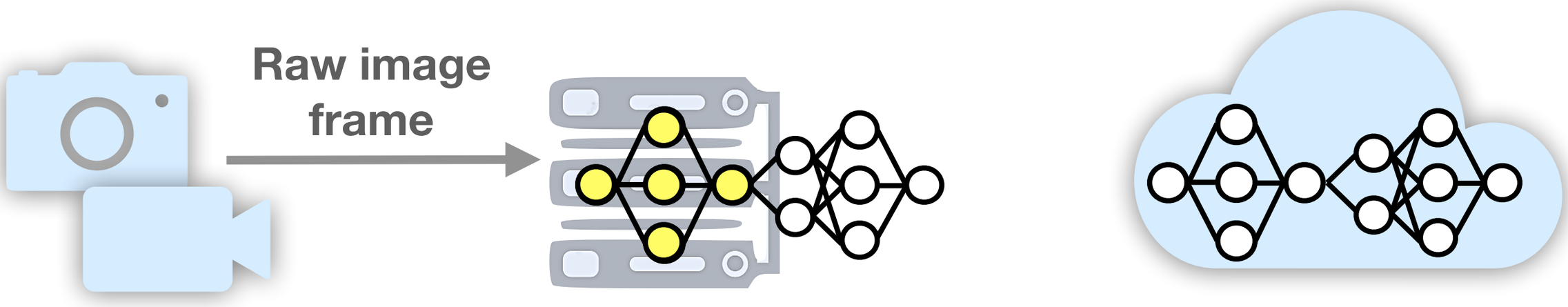
Couper: a general edge system

finding(and deploying) a good DNN deployment for you!

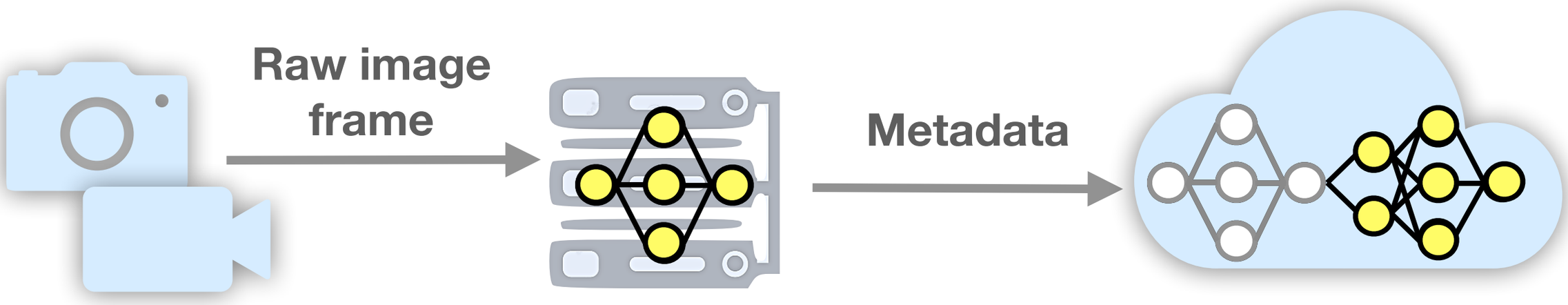
Share load across edge and cloud by DNN partitioning



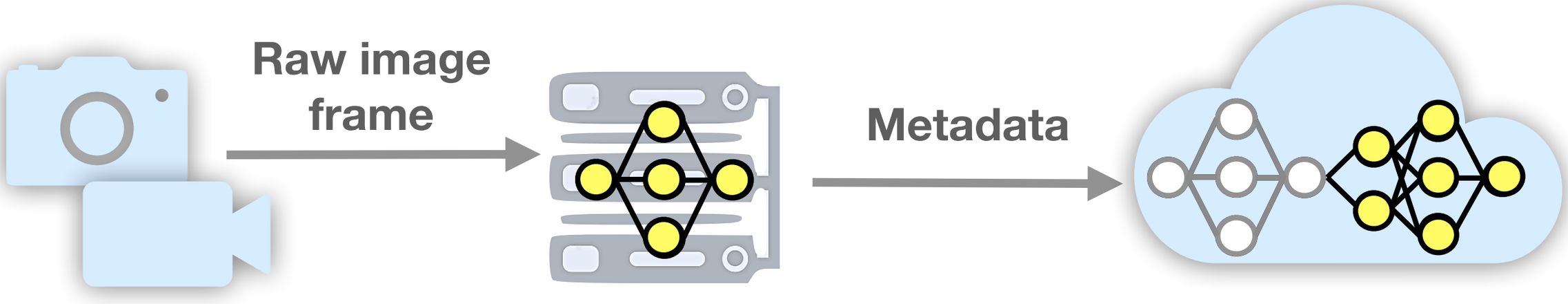
Share load across edge and cloud by DNN partitioning



Share load across edge and cloud by DNN partitioning

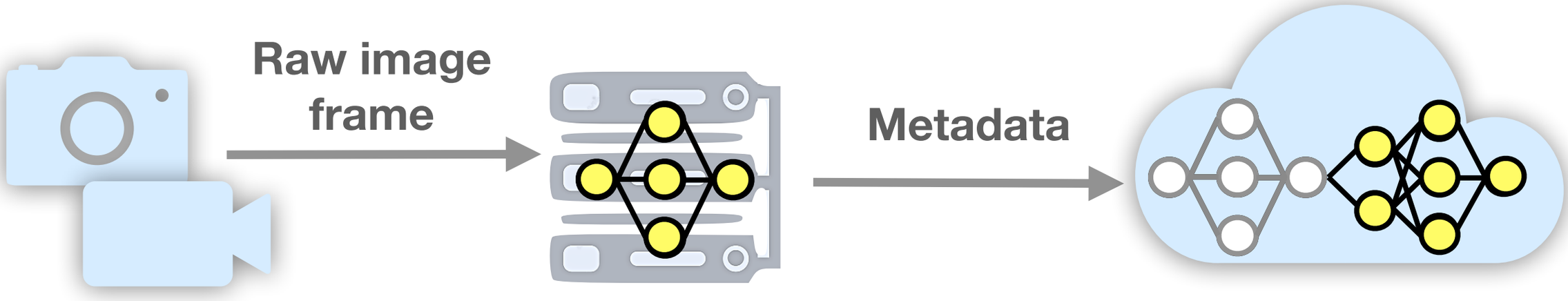


Share load across edge and cloud by DNN partitioning



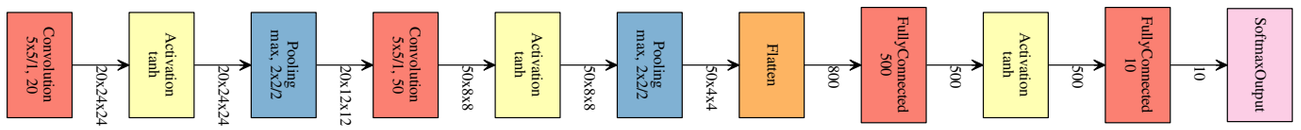
How do we decide the slicing point?

Share load across edge and cloud by DNN partitioning



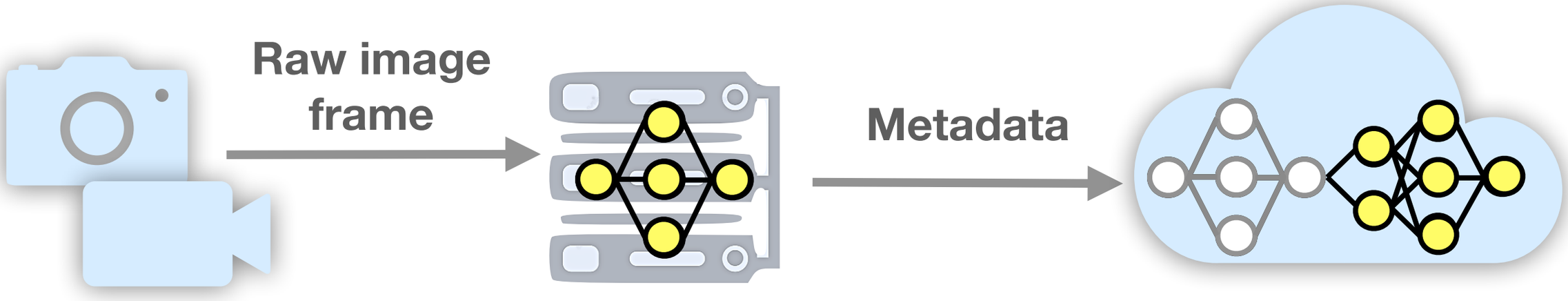
How do we decide the slicing point?

LeNet (1998)



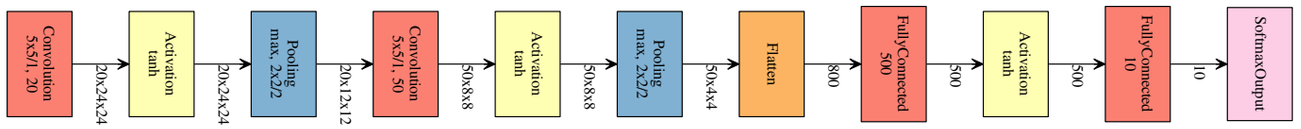
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning



How do we decide the slicing point?

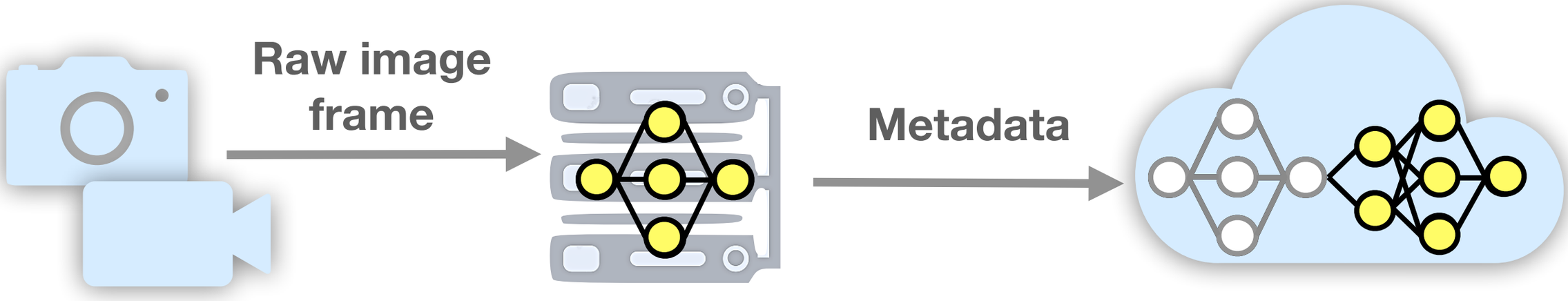
LeNet (1998)



AlexNet (2012)

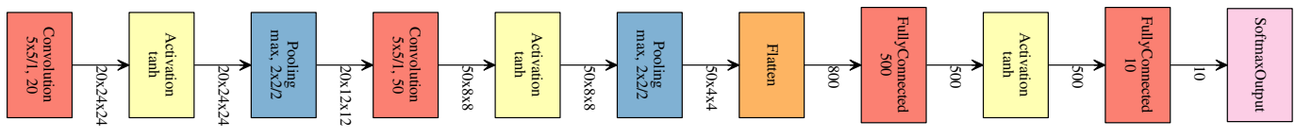
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning



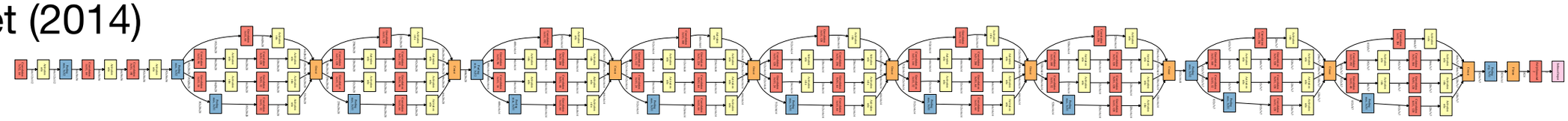
How do we decide the slicing point?

LeNet (1998)



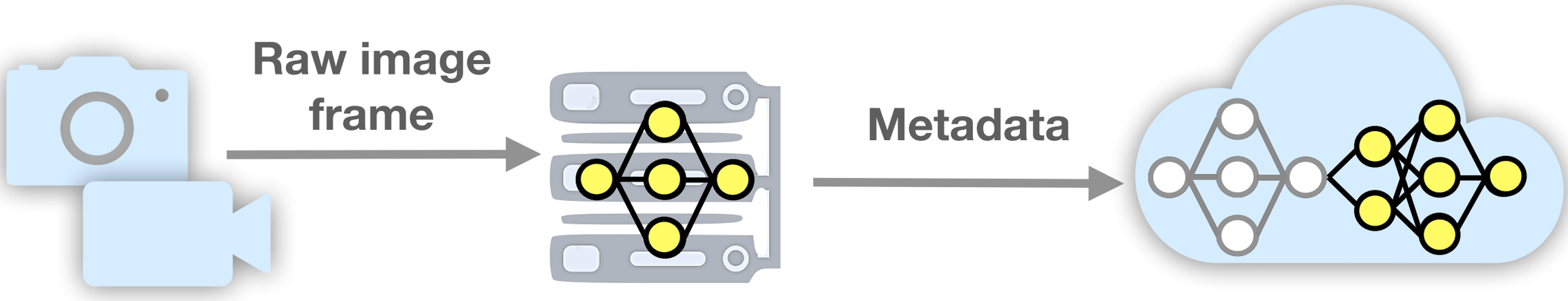
AlexNet (2012)

GoogLeNet (2014)



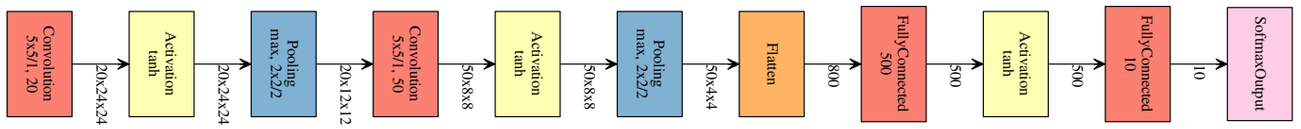
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning



How do we decide the slicing point?

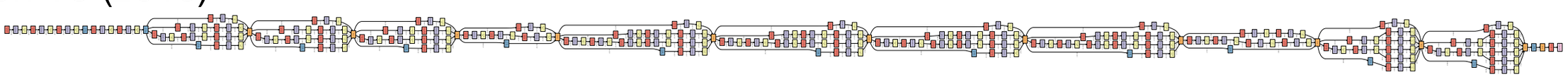
LeNet (1998)



AlexNet (2012)

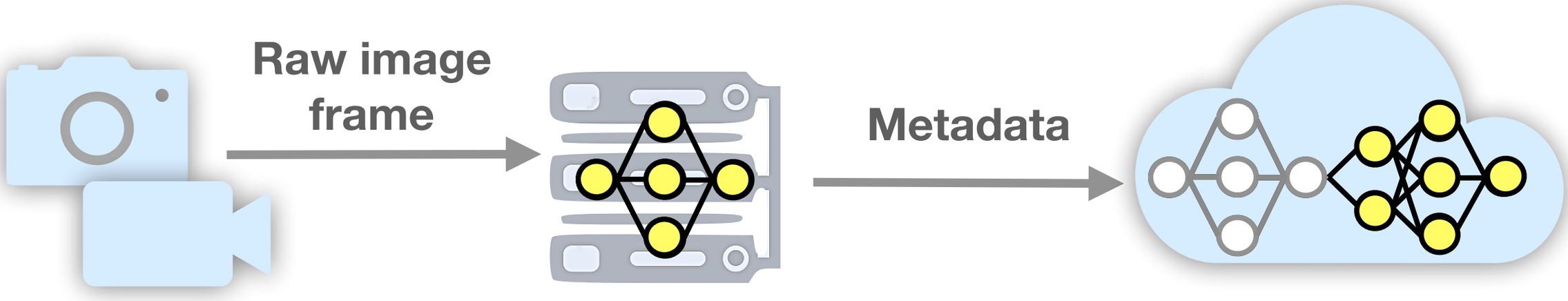
GoogLeNet (2014)

Inception V3 (2015)



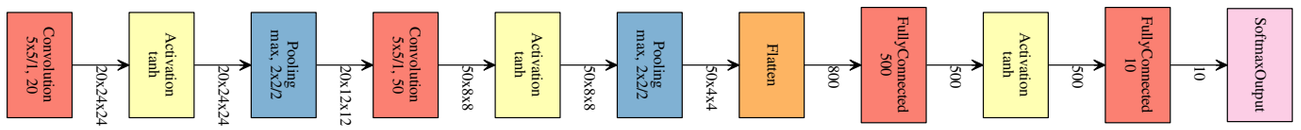
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning



How do we decide the slicing point?

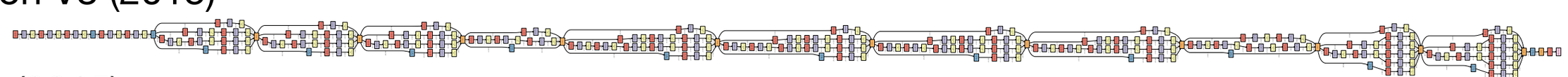
LeNet (1998)



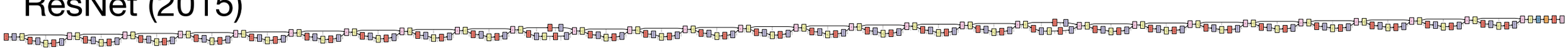
AlexNet (2012)

GoogLeNet (2014)

Inception V3 (2015)

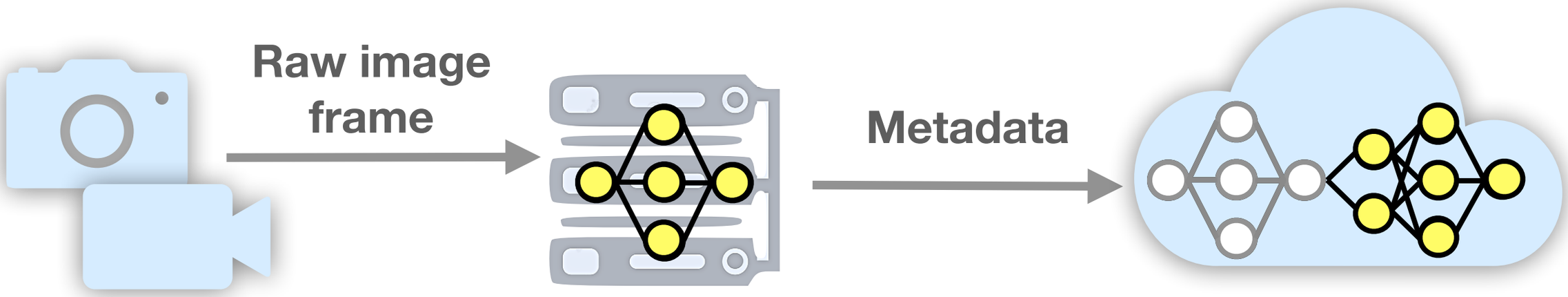


ResNet (2015)



http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning

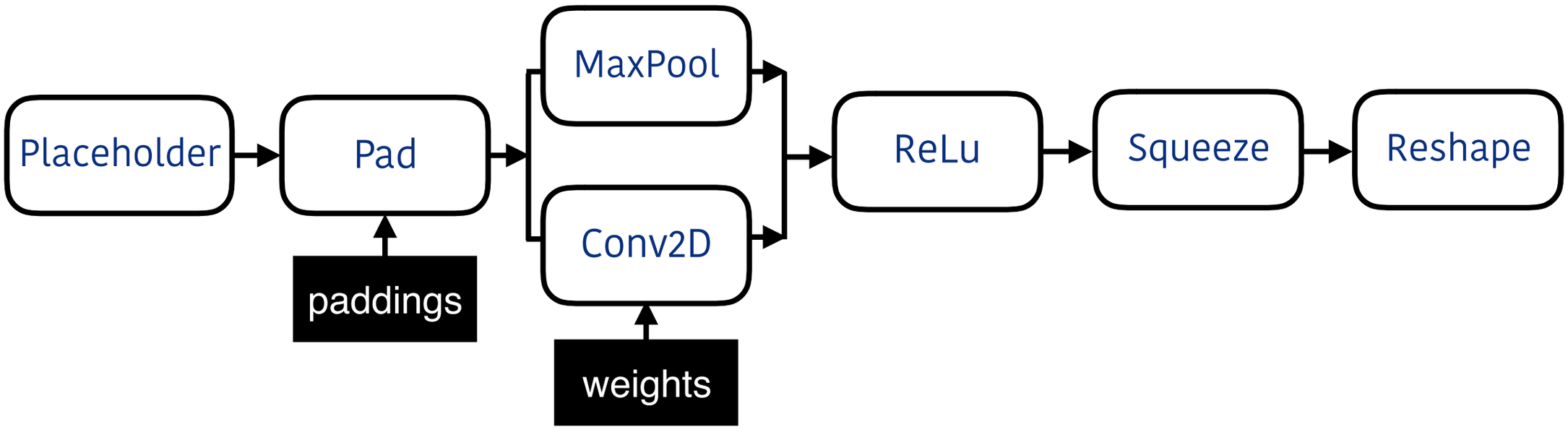


How do we decide the partition point?

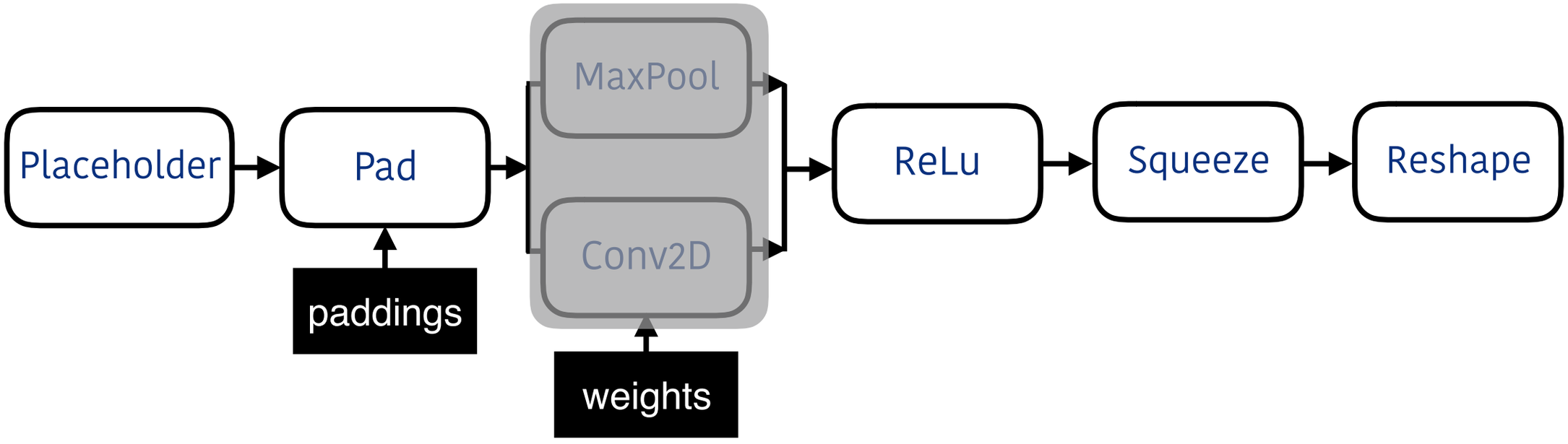
1. Filter out splittable candidates in DNN

2. Pick up a right one among the candidates

Listing splicing candidates

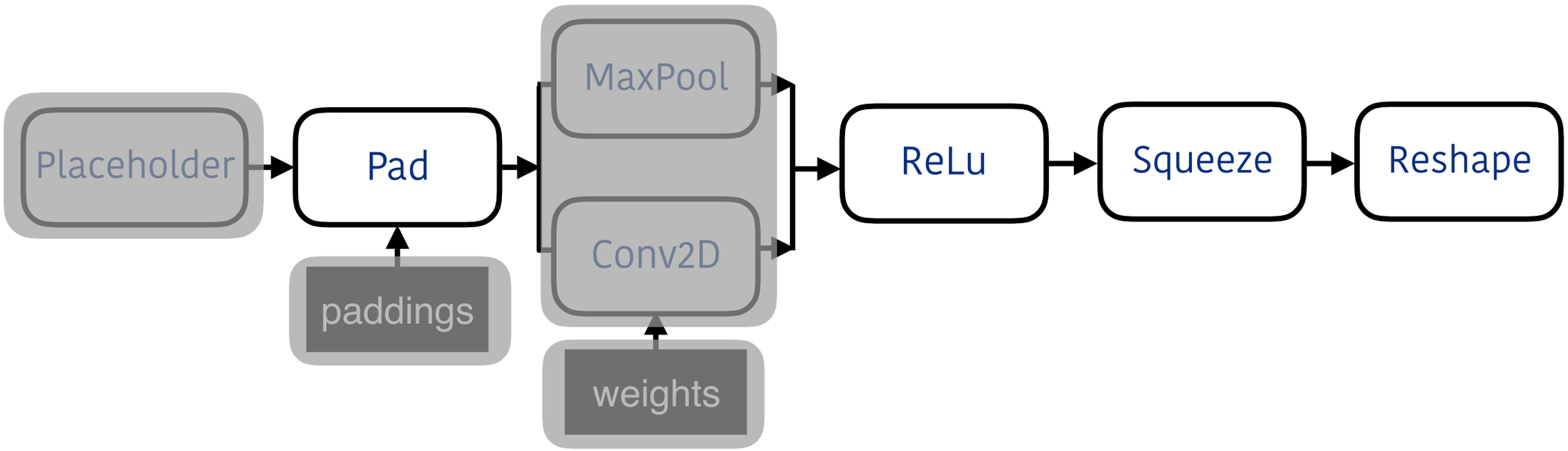


Listing splicing candidates



✗ Multi-parallel path

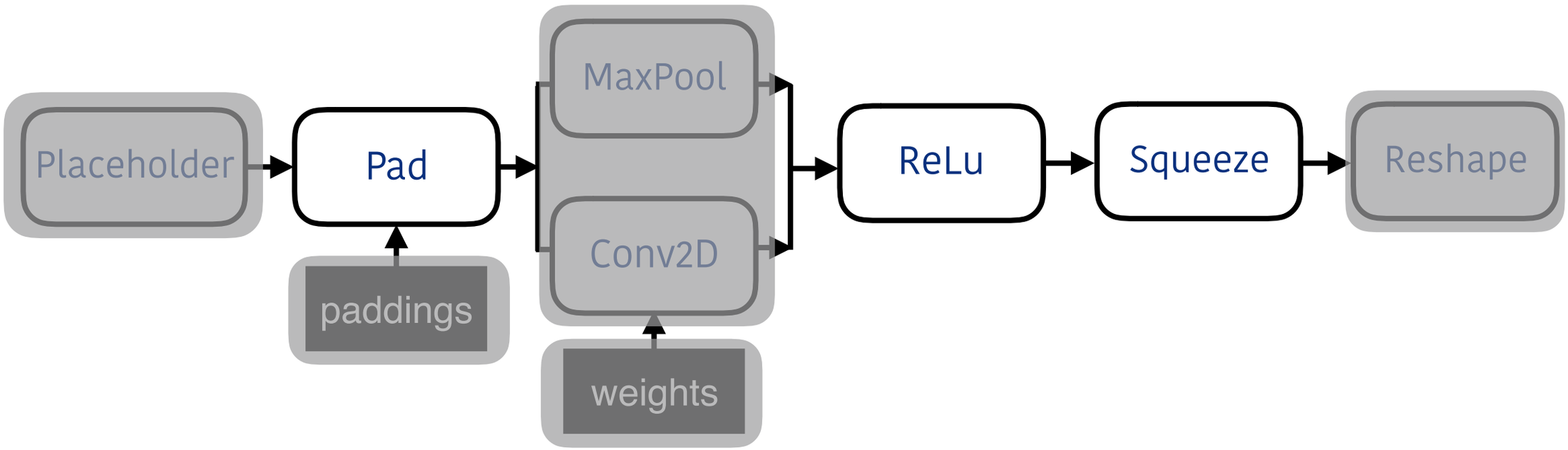
Listing splicing candidates



✗ Multi-parallel path

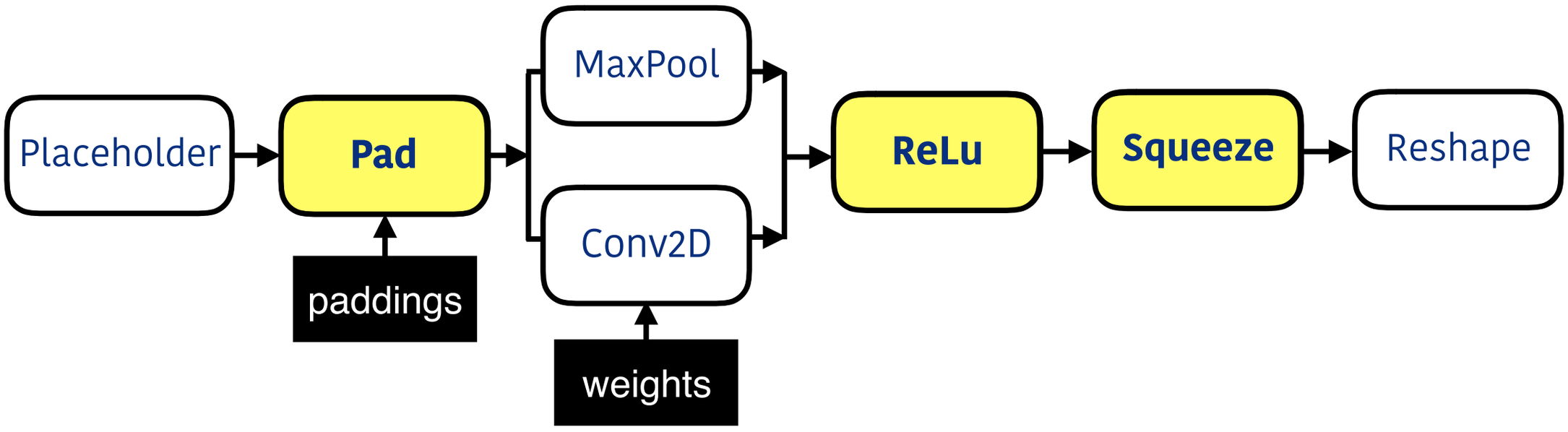
✗ Constant or reading operator

Listing splicing candidates



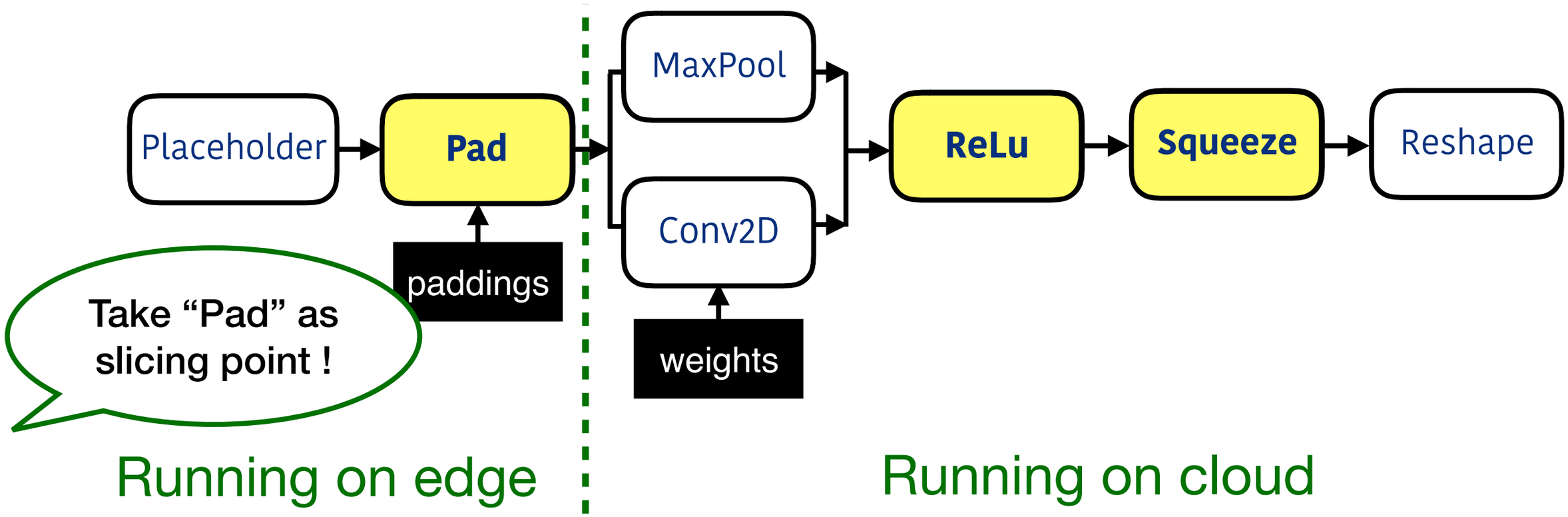
- ✗ Multi-parallel path
- ✗ Constant or reading operator
- ✗ Last operator

Listing splicing candidates



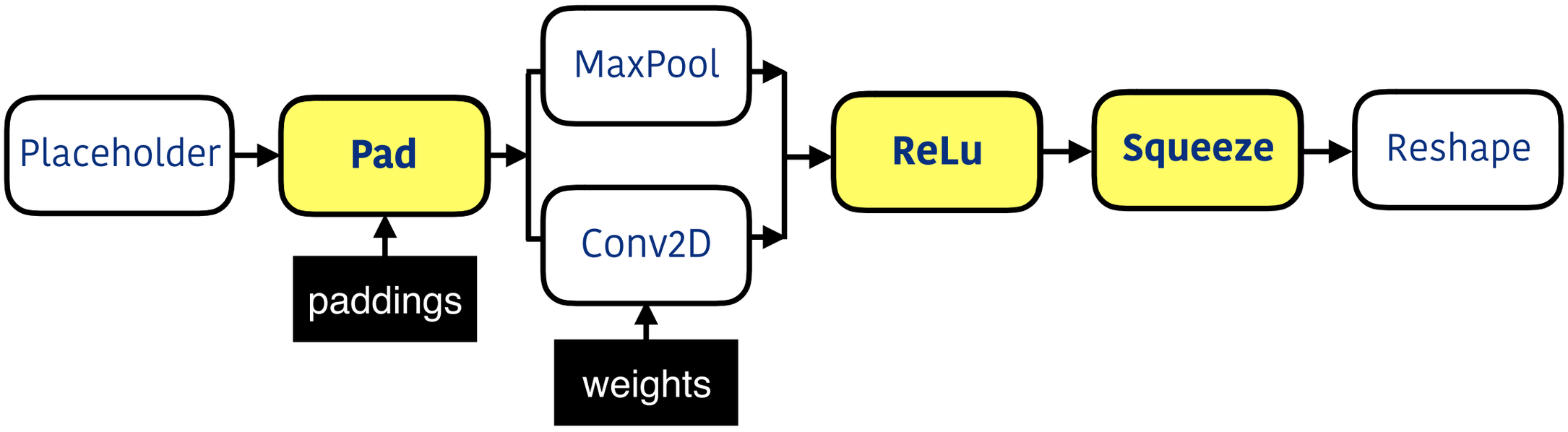
- ✗ Multi-parallel path
- ✗ Constant or reading operator
- ✗ Last operator

Listing splicing candidates

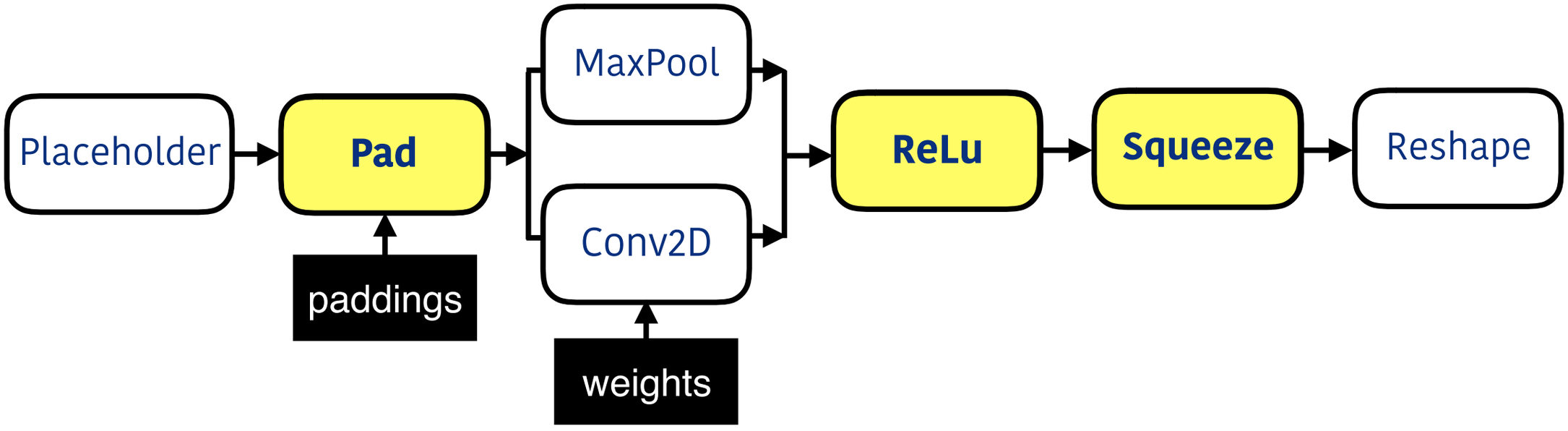


- ✗ Multi-parallel path
- ✗ Constant or reading operator
- ✗ Last operator

Evaluating splicing candidates

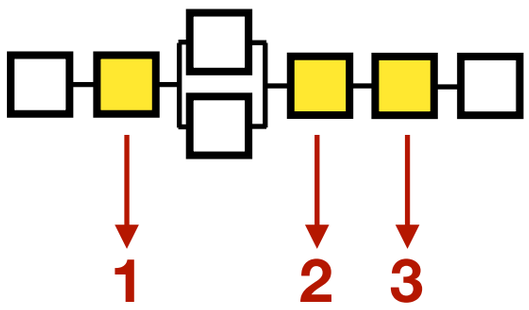


Evaluating splicing candidates

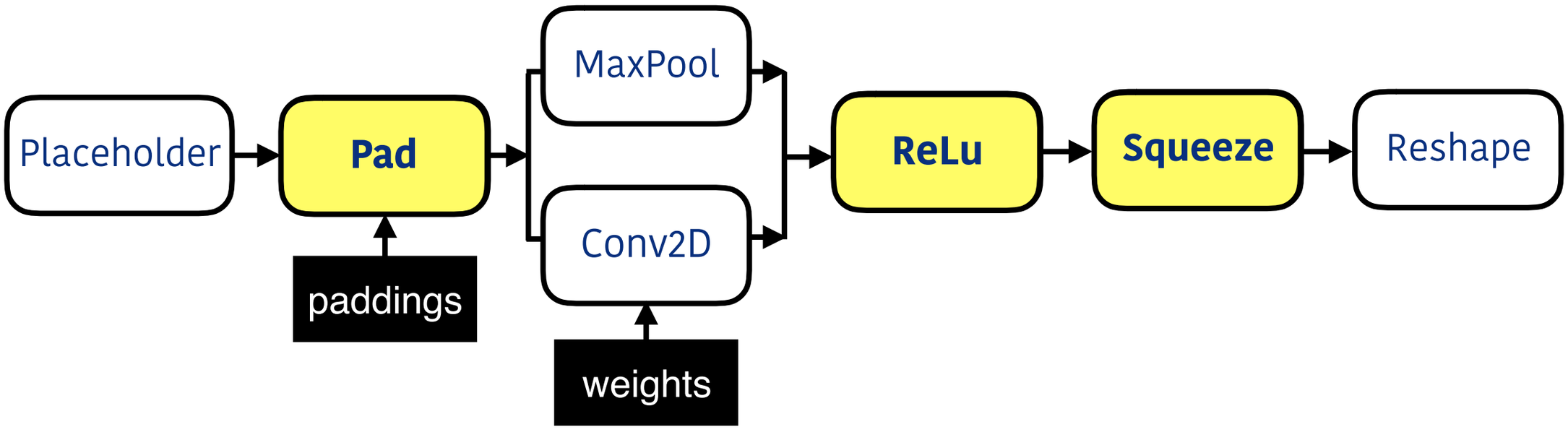


Strongman

Evaluate
every candidate

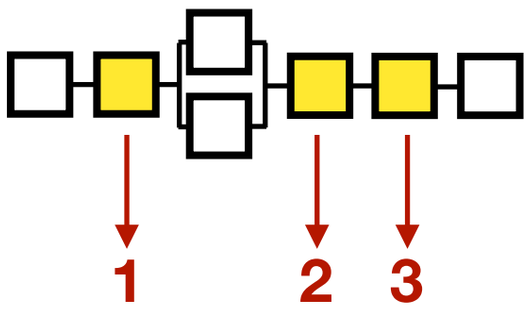


Evaluating splicing candidates



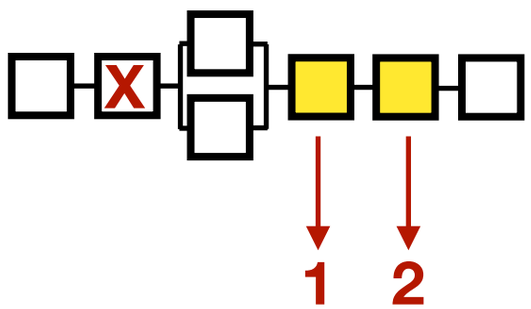
Strongman

Evaluate
every candidate

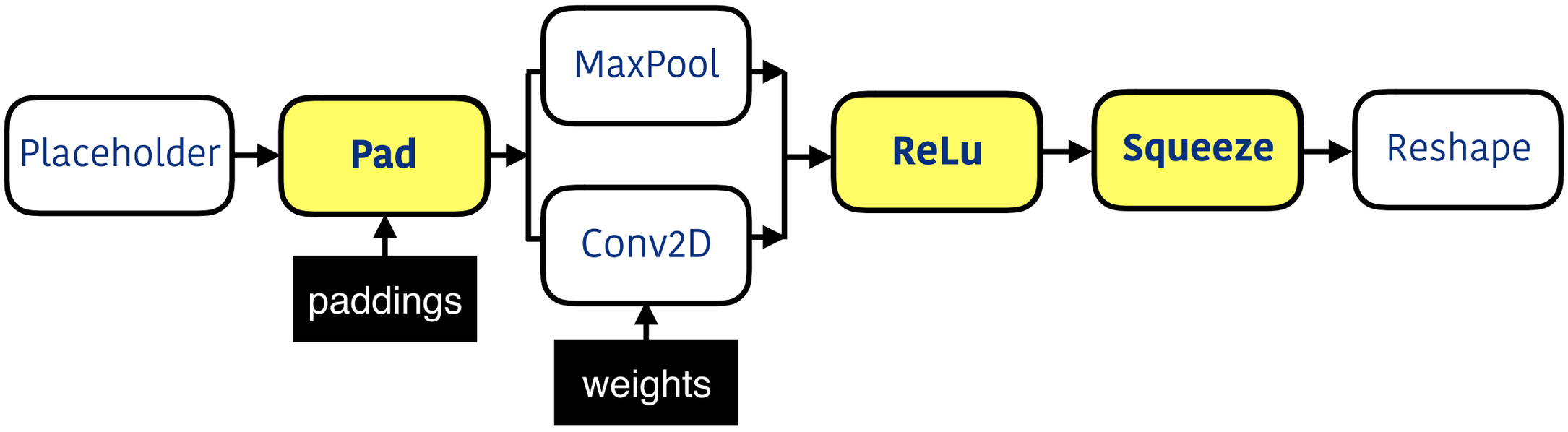


Comm-slim

Bypass candidates with
high networking cost

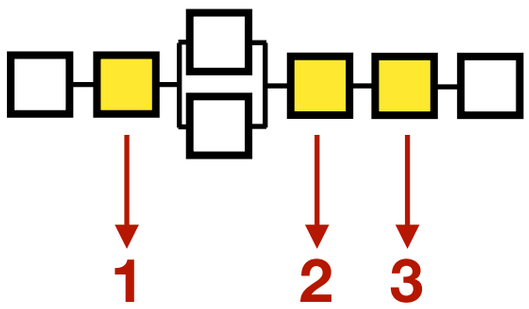


Evaluating splicing candidates



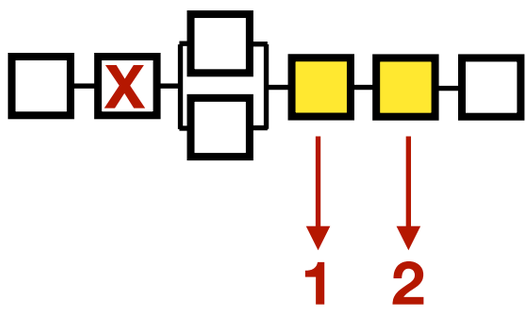
Strongman

Evaluate
every candidate



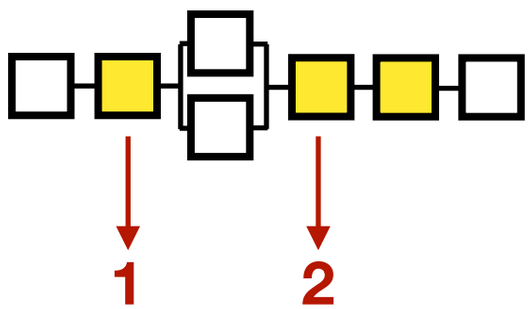
Comm-slim

Bypass candidates with
high networking cost

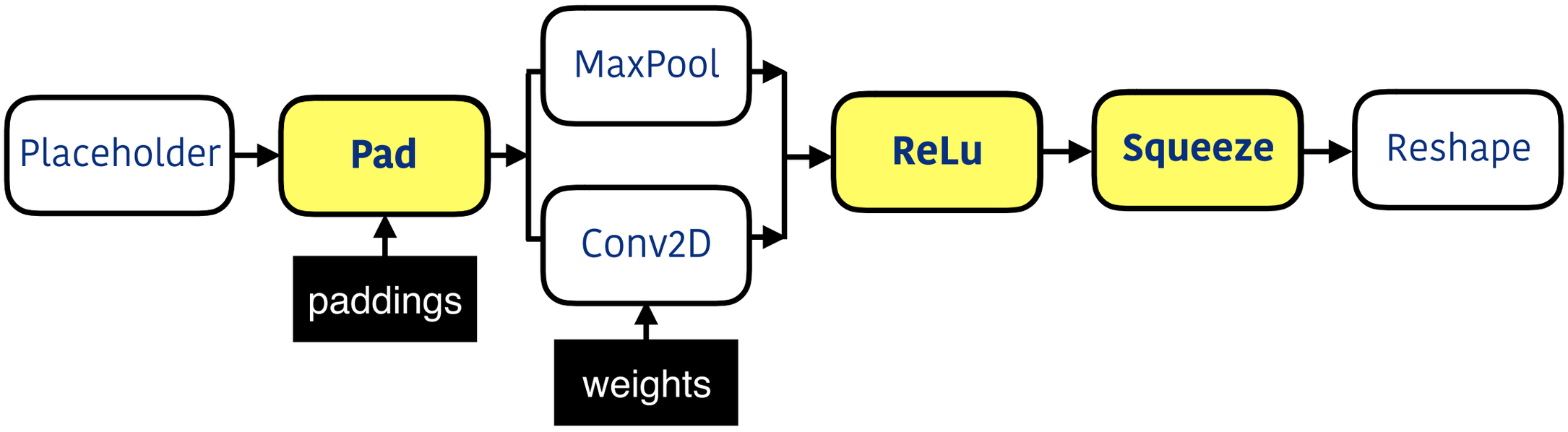


Early-stop

Stop evaluation
when edge is overload

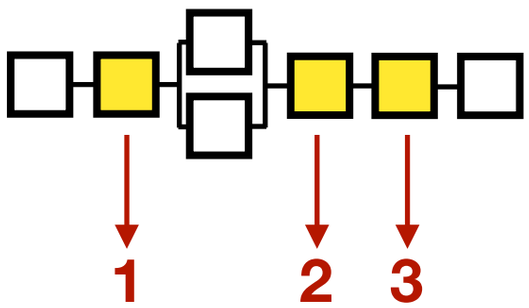


Evaluating splicing candidates



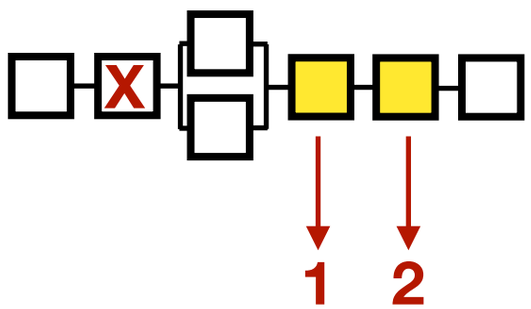
Strongman

Evaluate every candidate



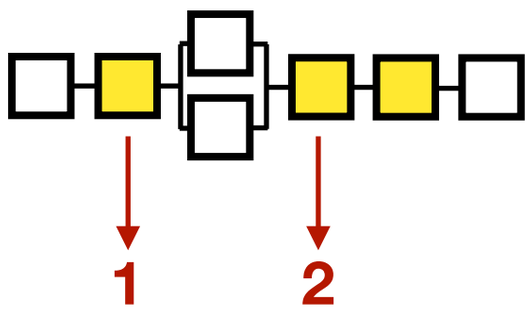
Comm-slim

Bypass candidates with high networking cost



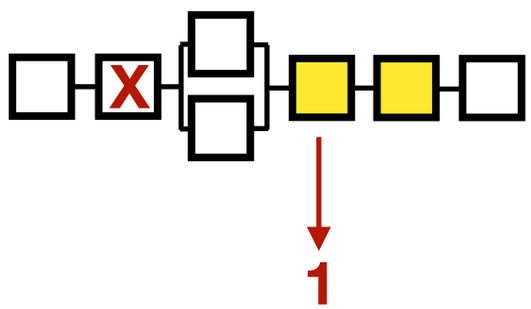
Early-stop

Stop evaluation when edge is overload

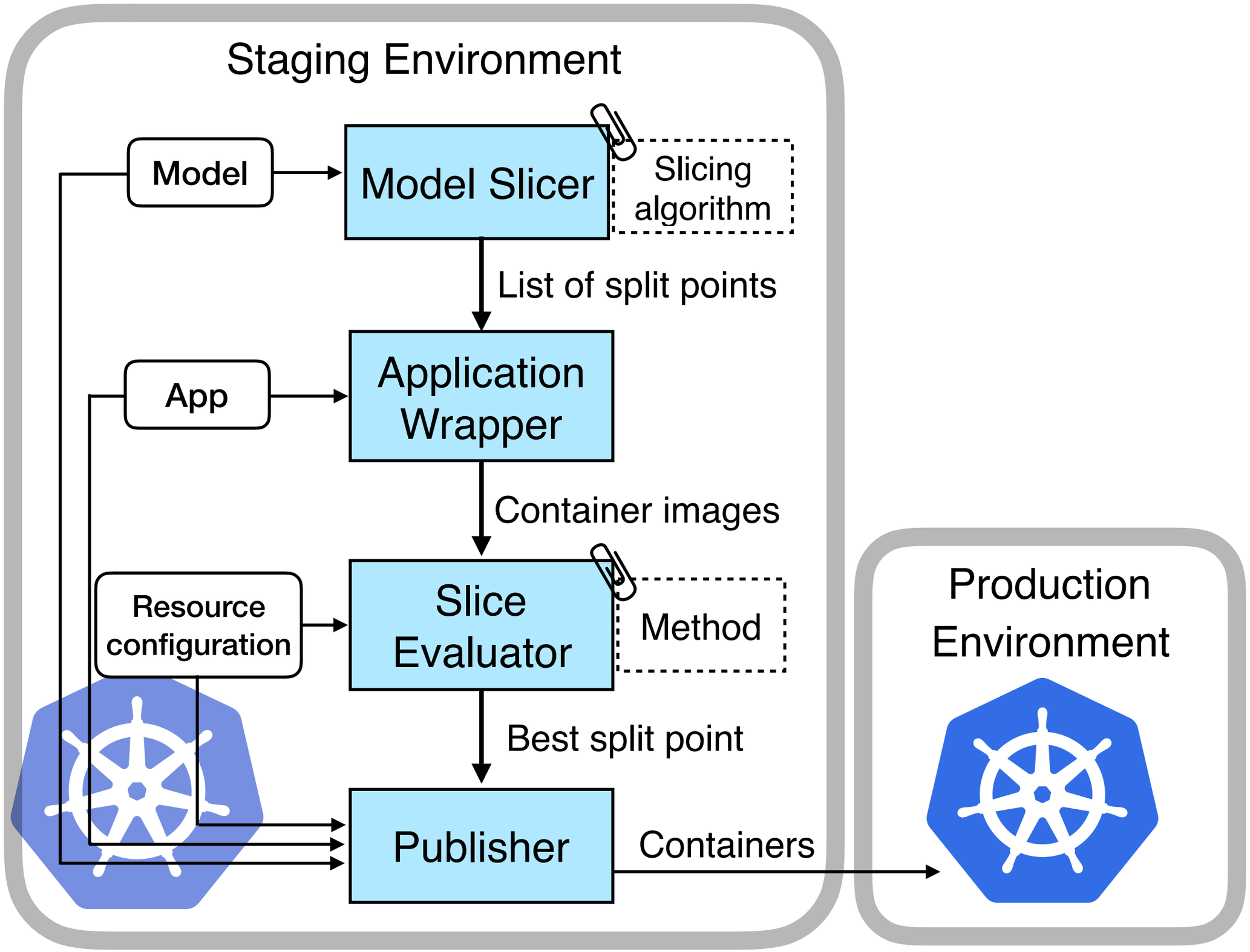


Hybrid

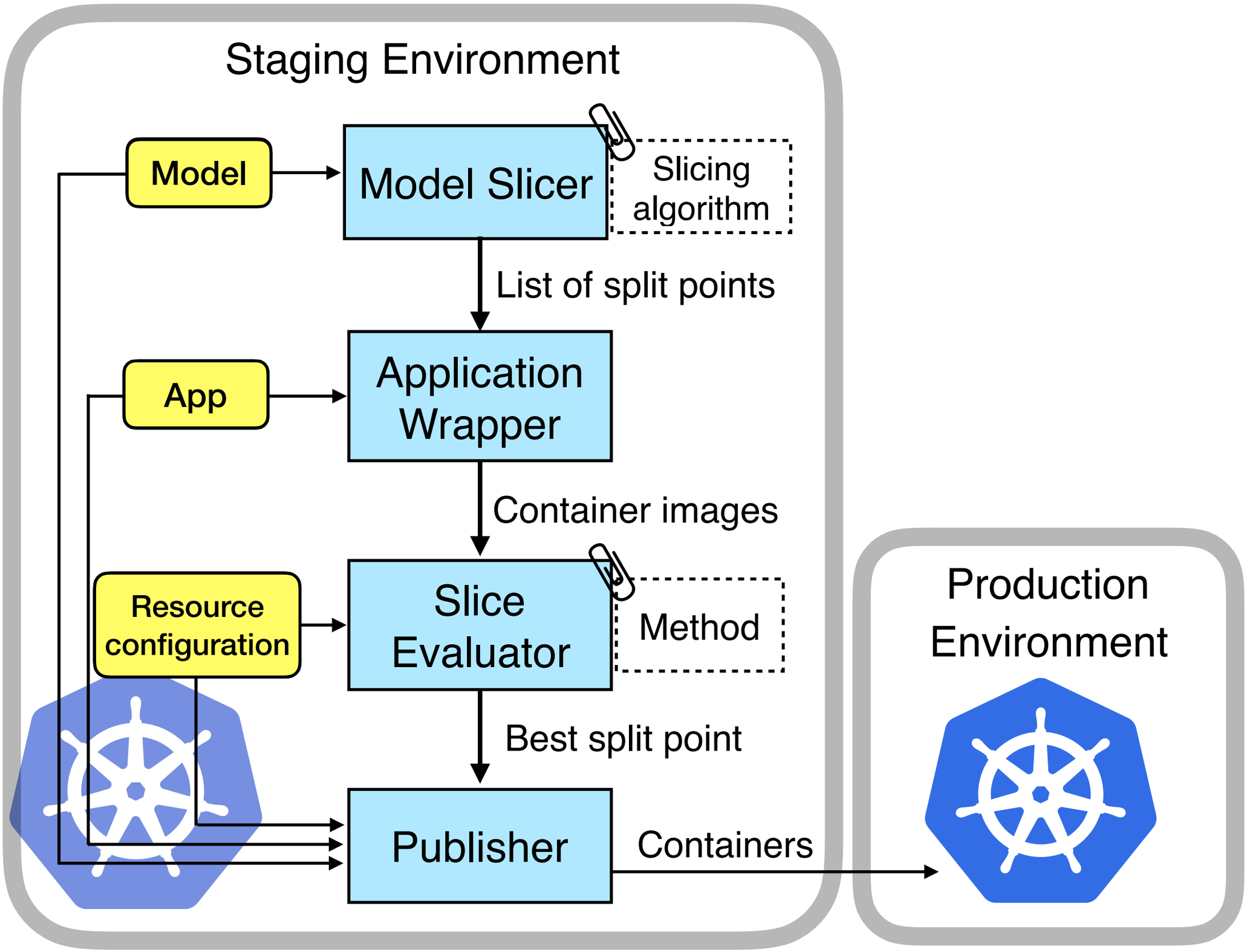
Combination of comm-slim and early-stop



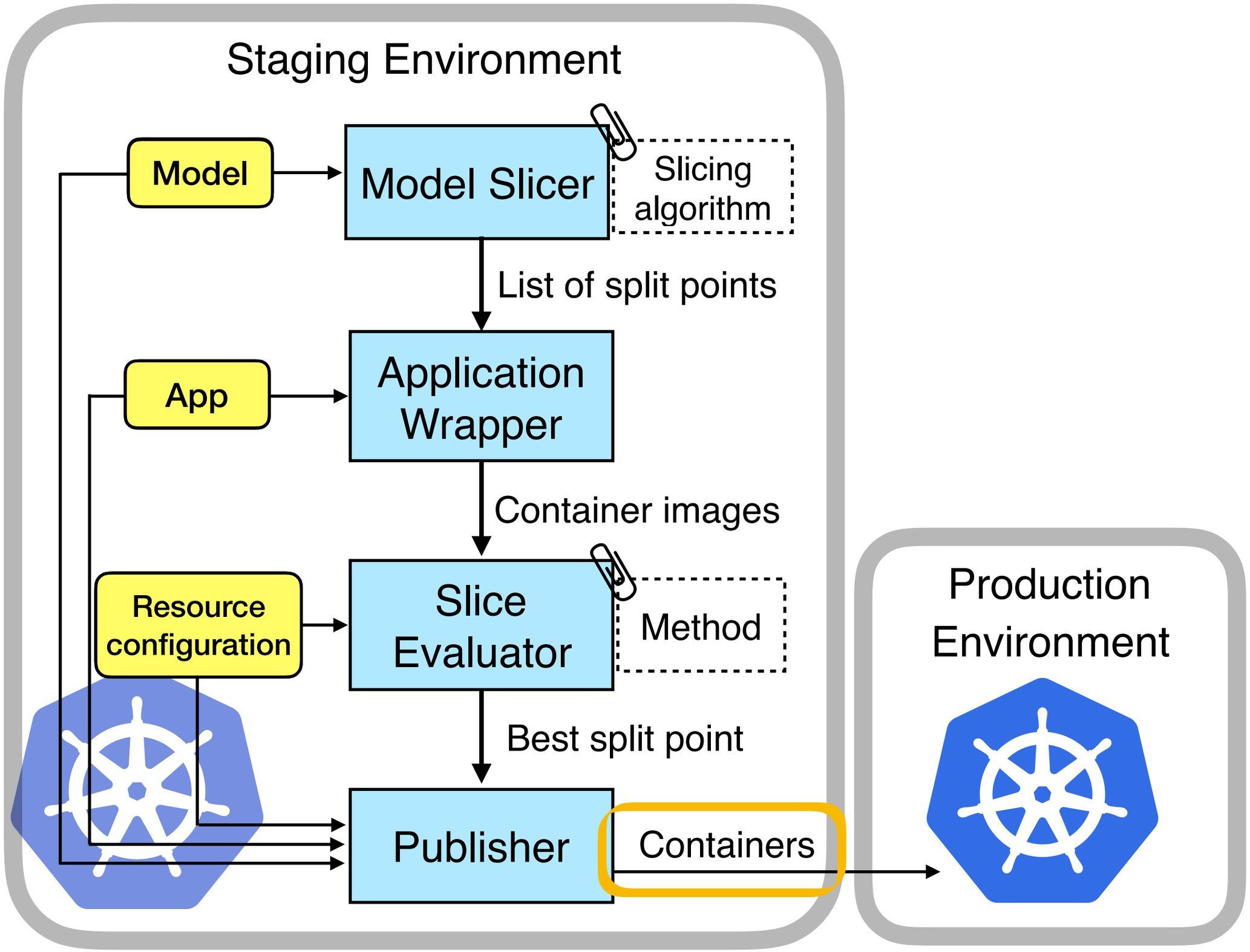
Couper Overview



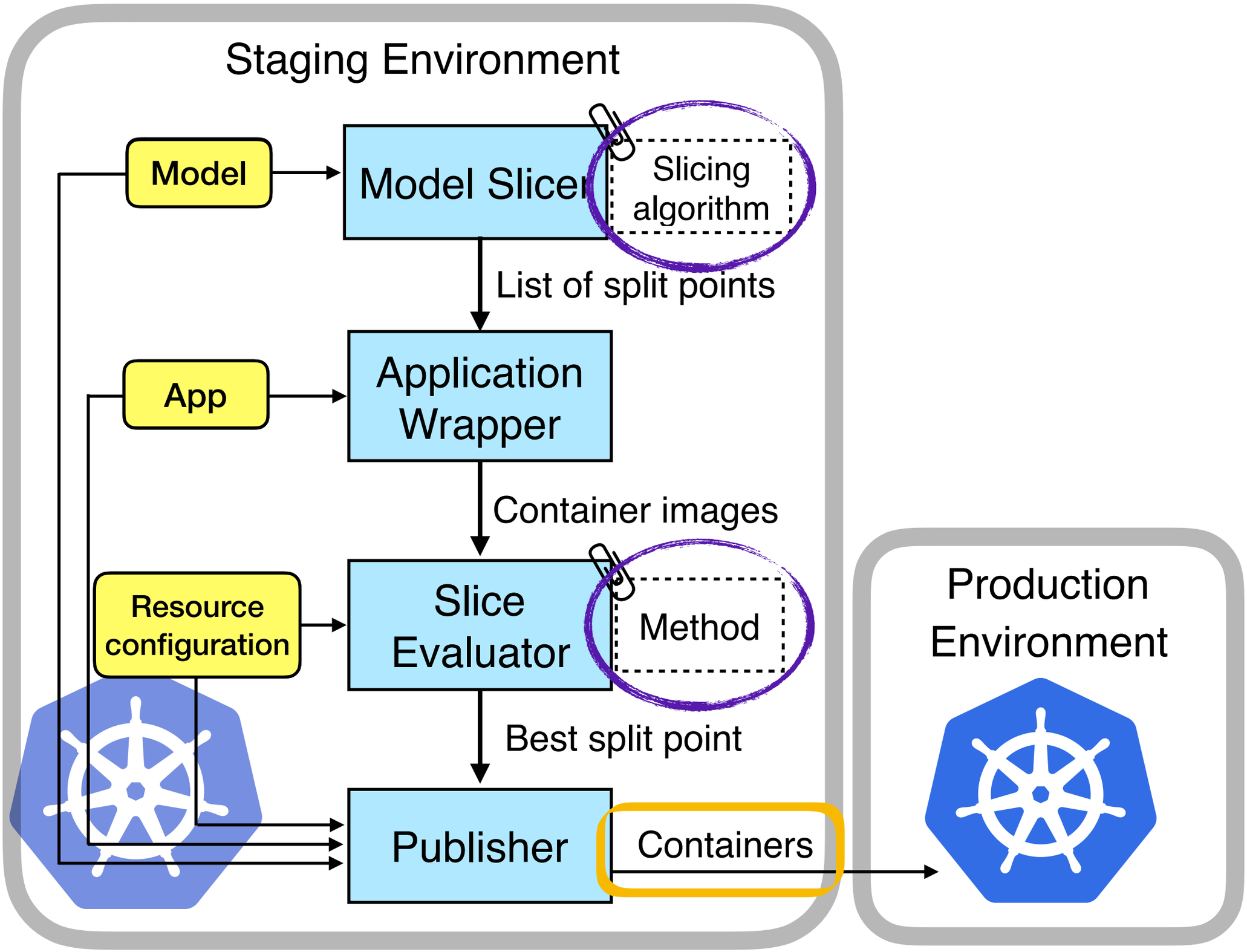
Couper Overview



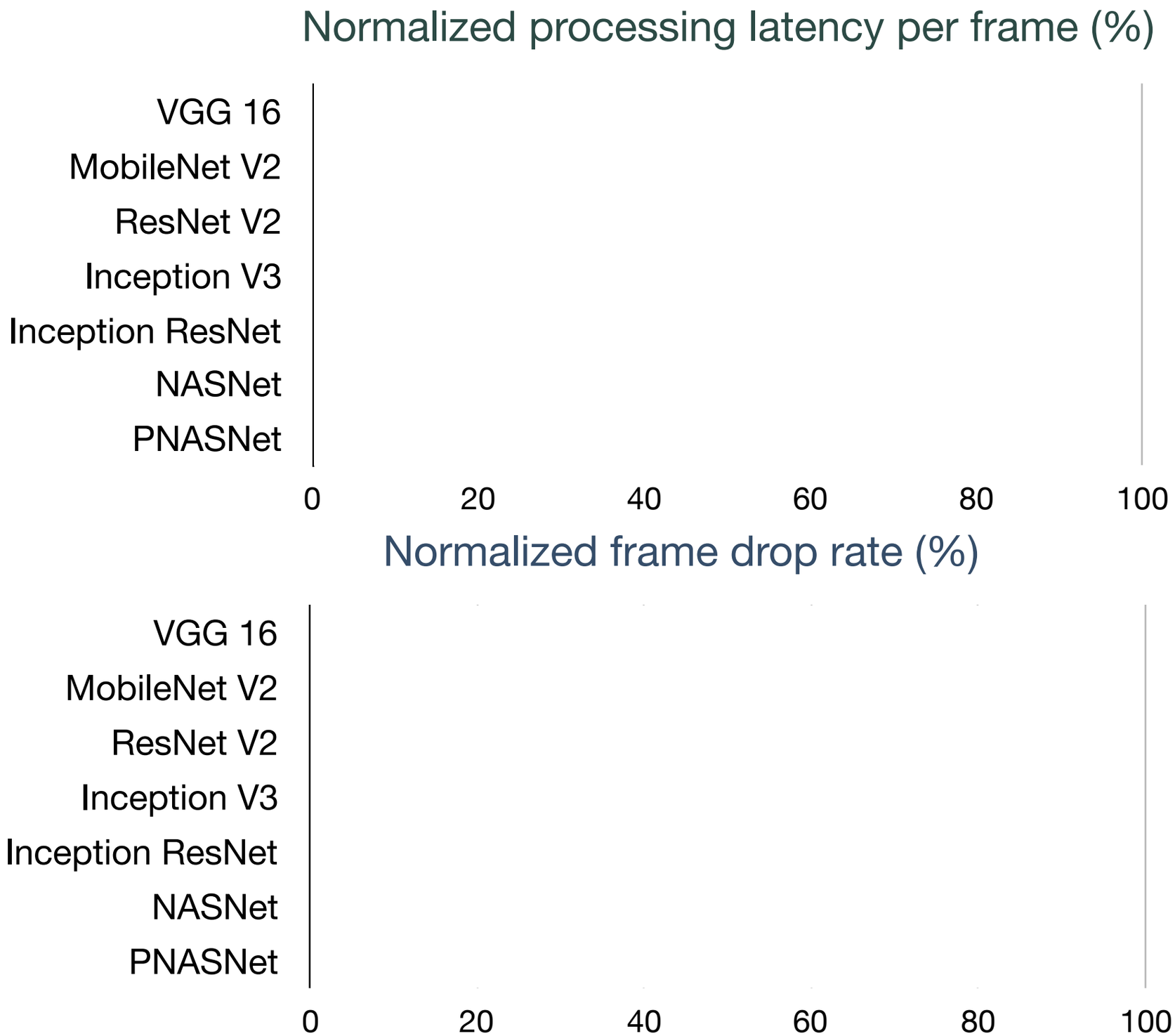
Couper Overview



Couper Overview

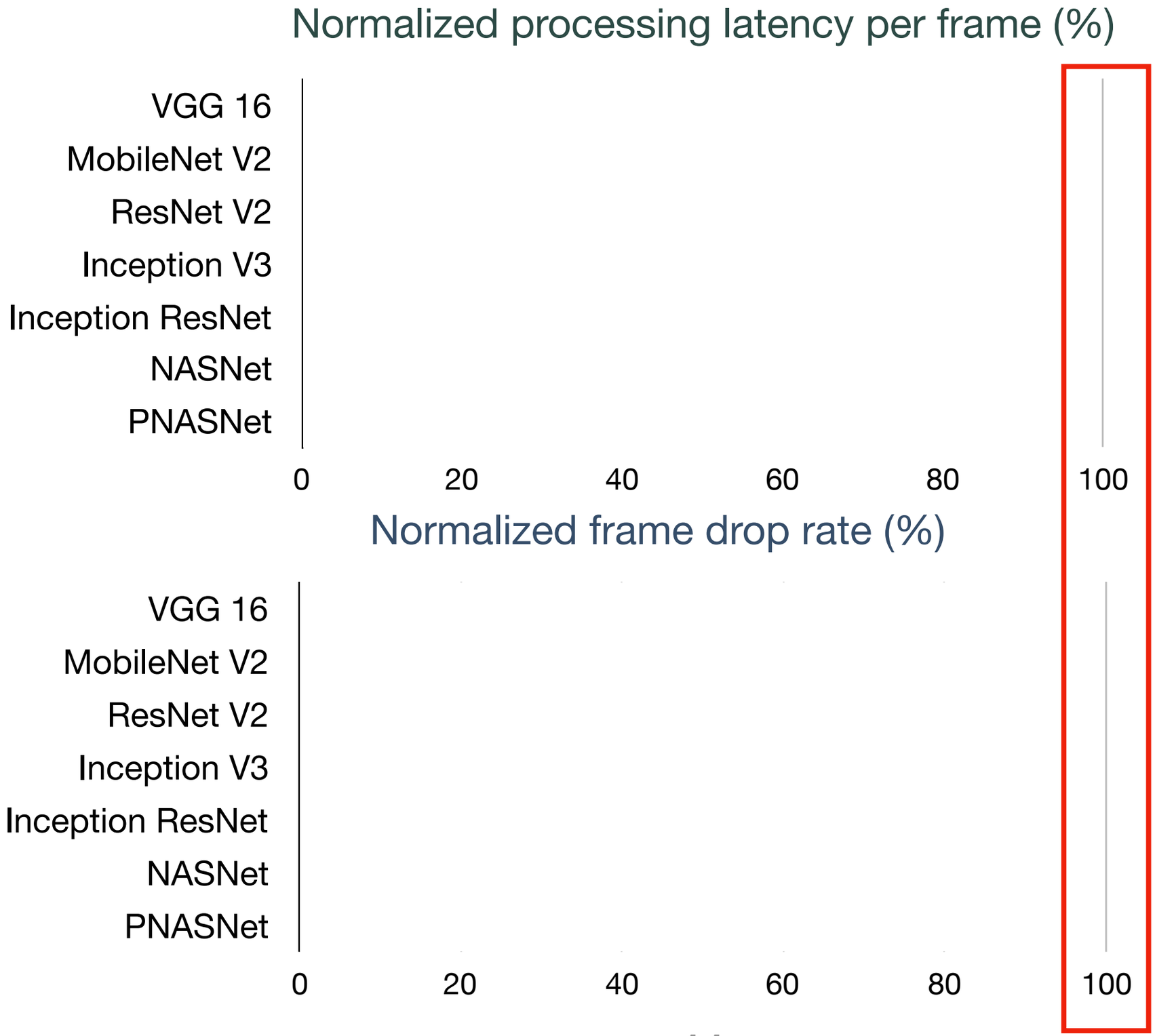


Results for different SLAs

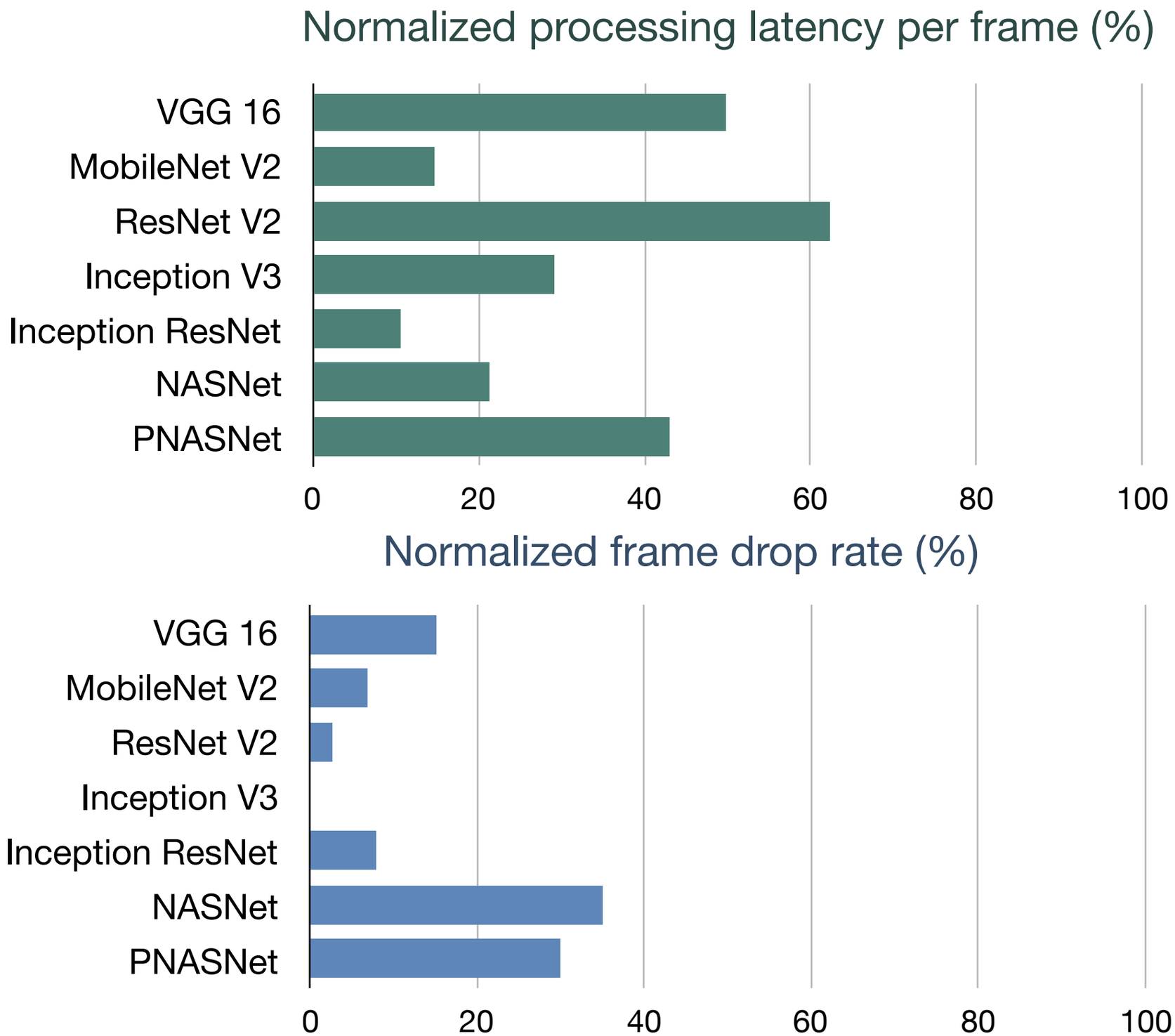


Results for different SLAs

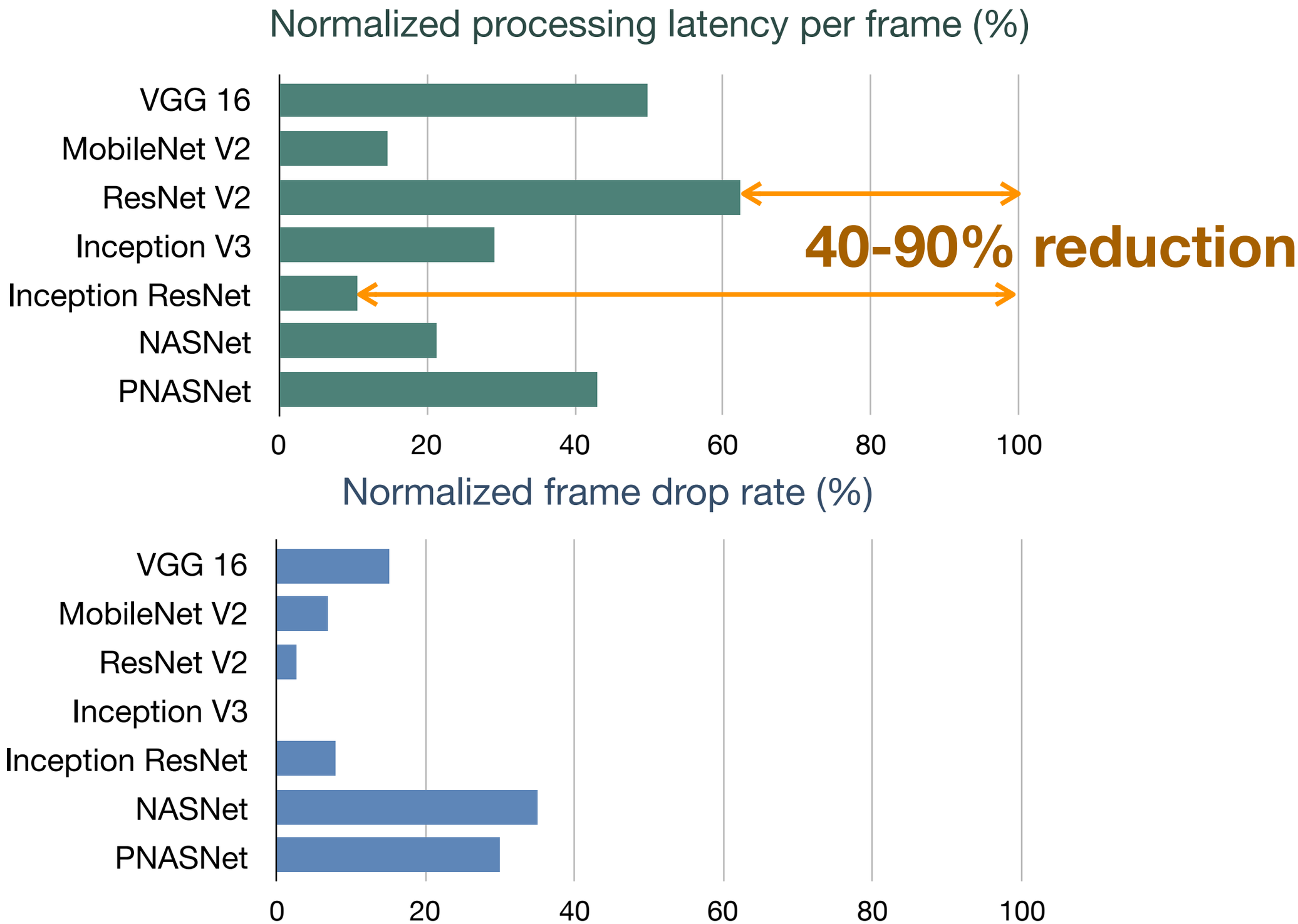
placing all DNN inference on cloud



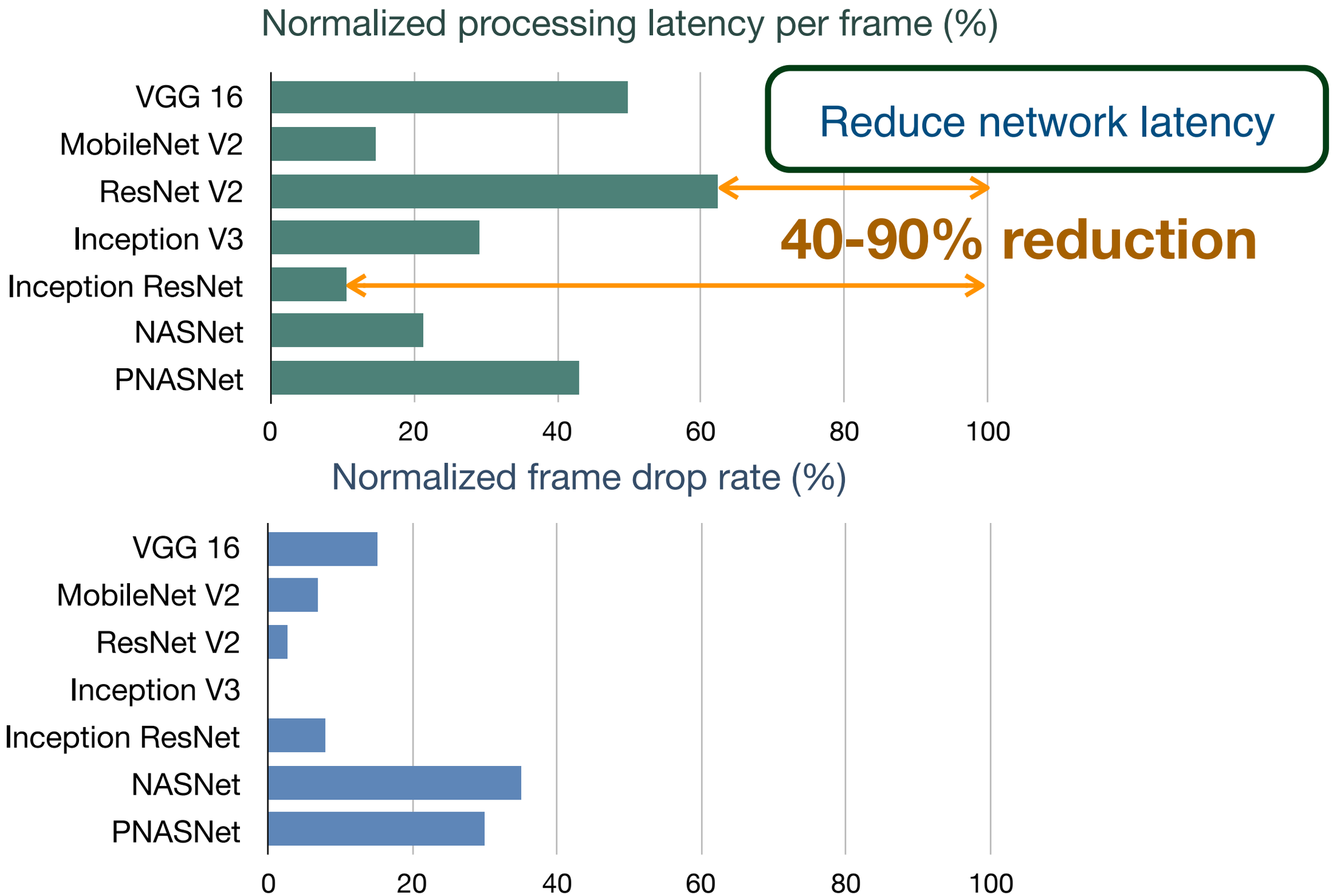
Results for different SLAs



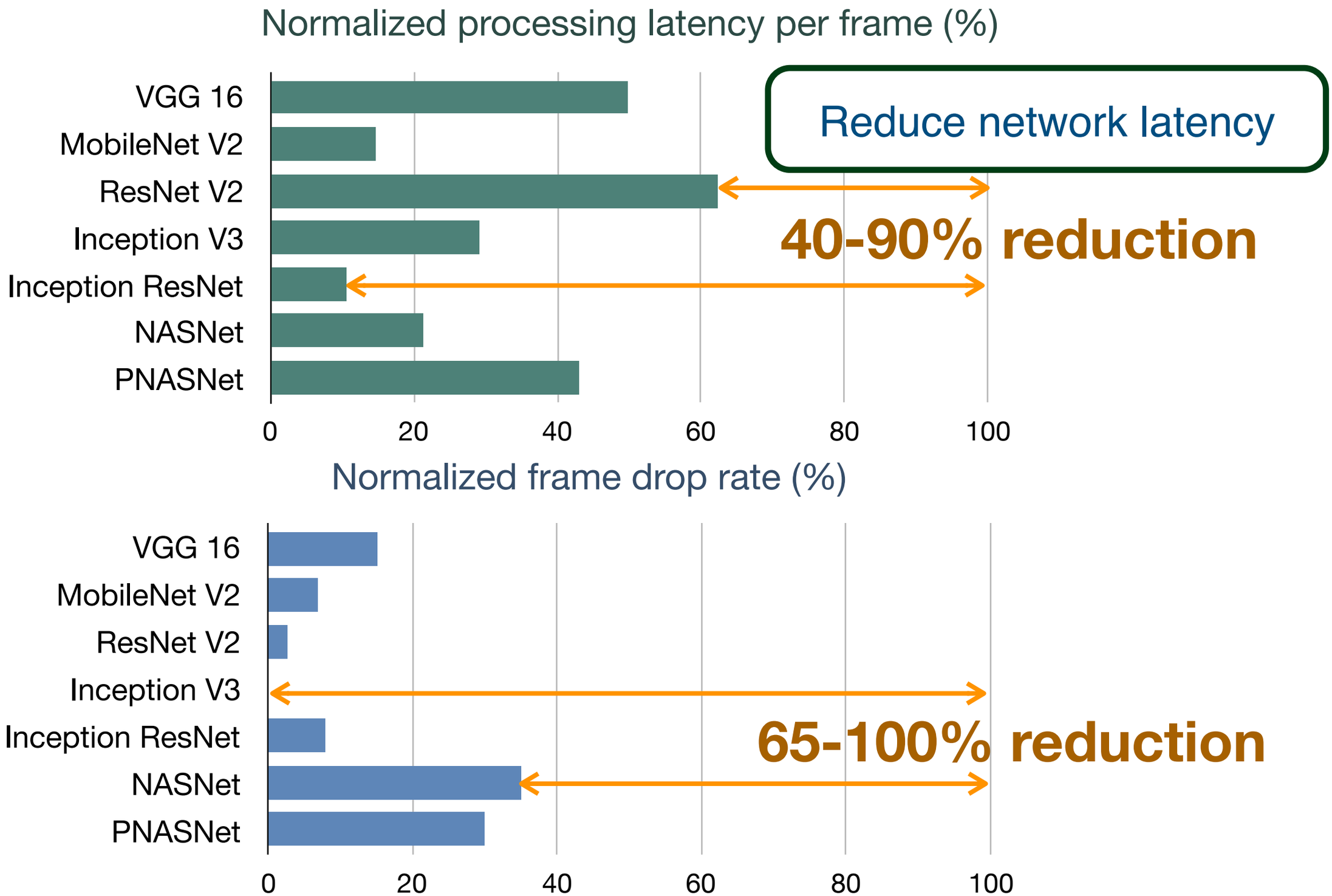
Results for different SLAs



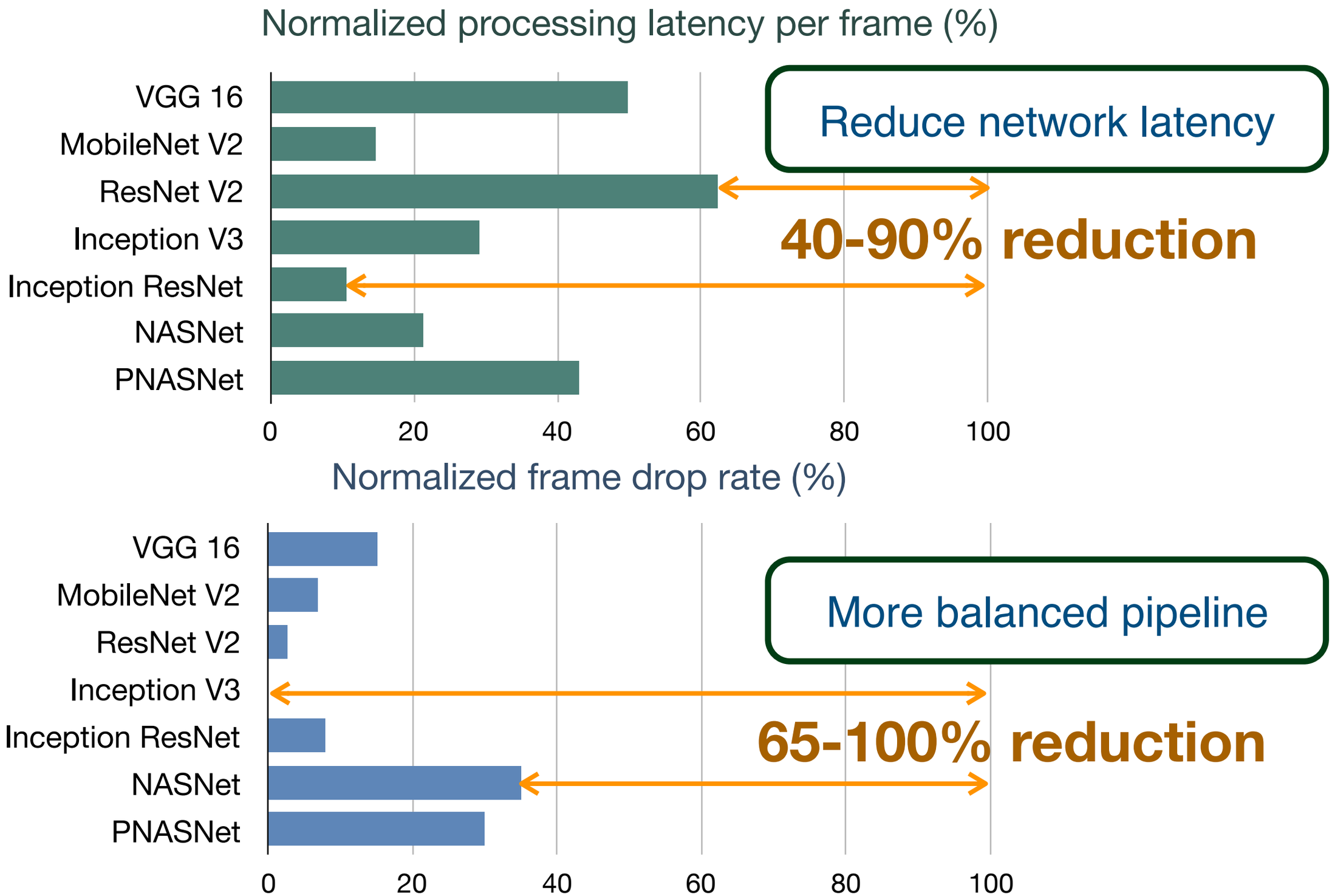
Results for different SLAs



Results for different SLAs



Results for different SLAs



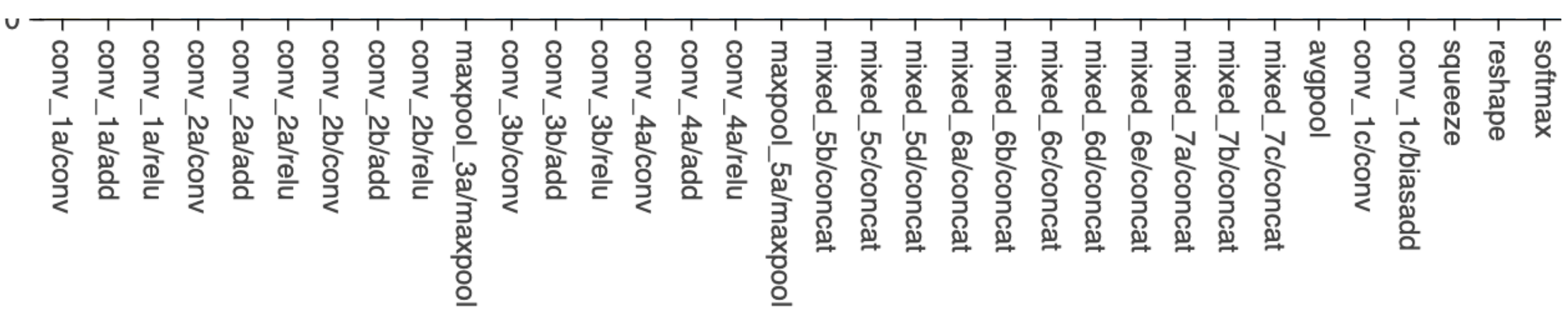
Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2

Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2

99% reduction

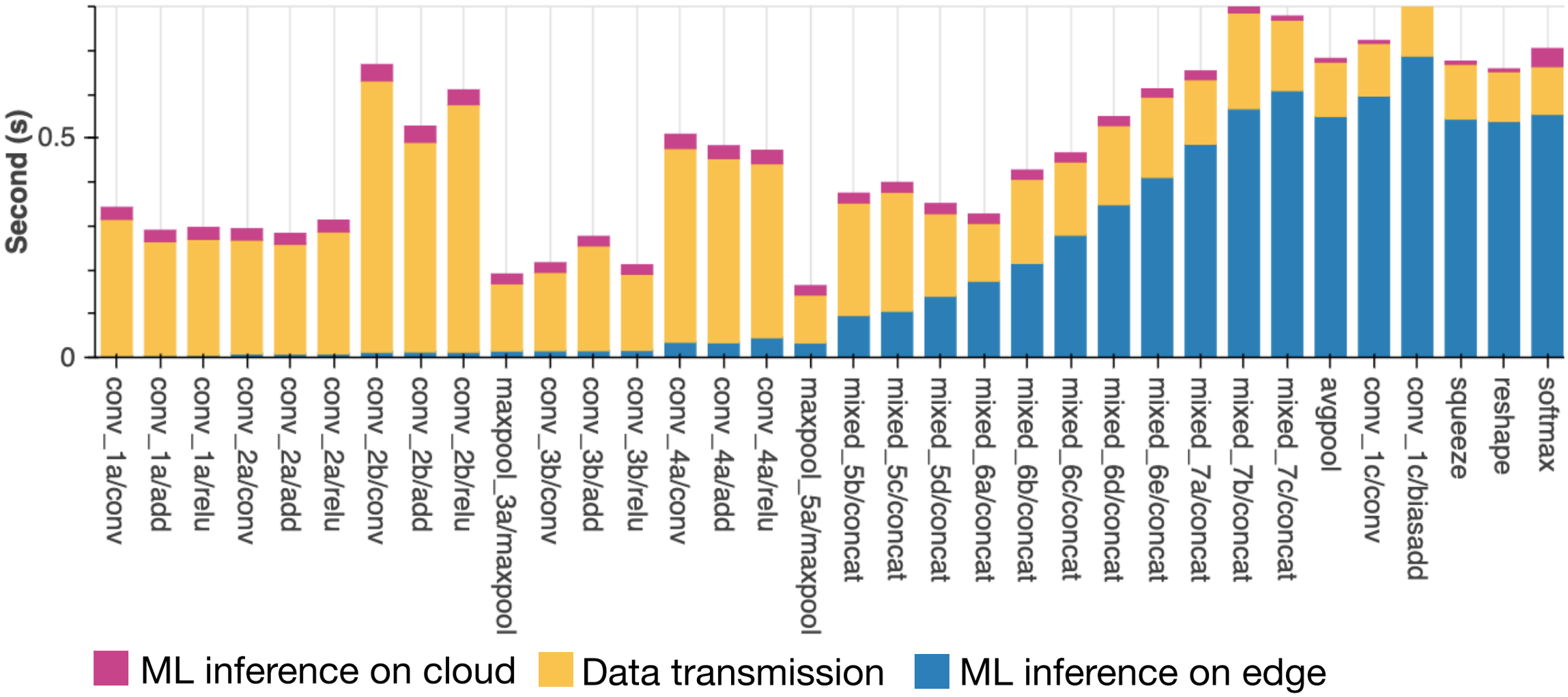
Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2

99% reduction

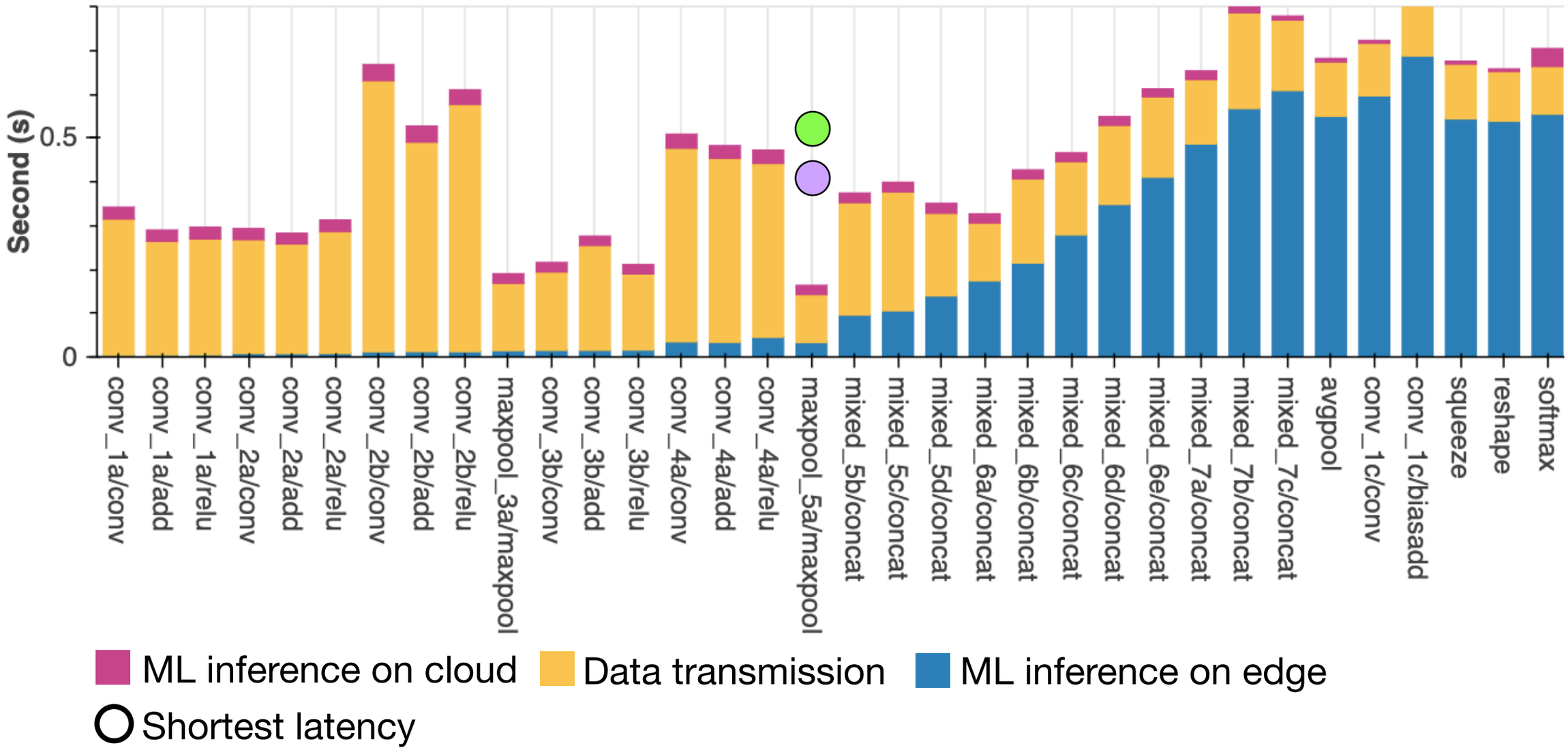


Strongman method tests 34 slicing candidates

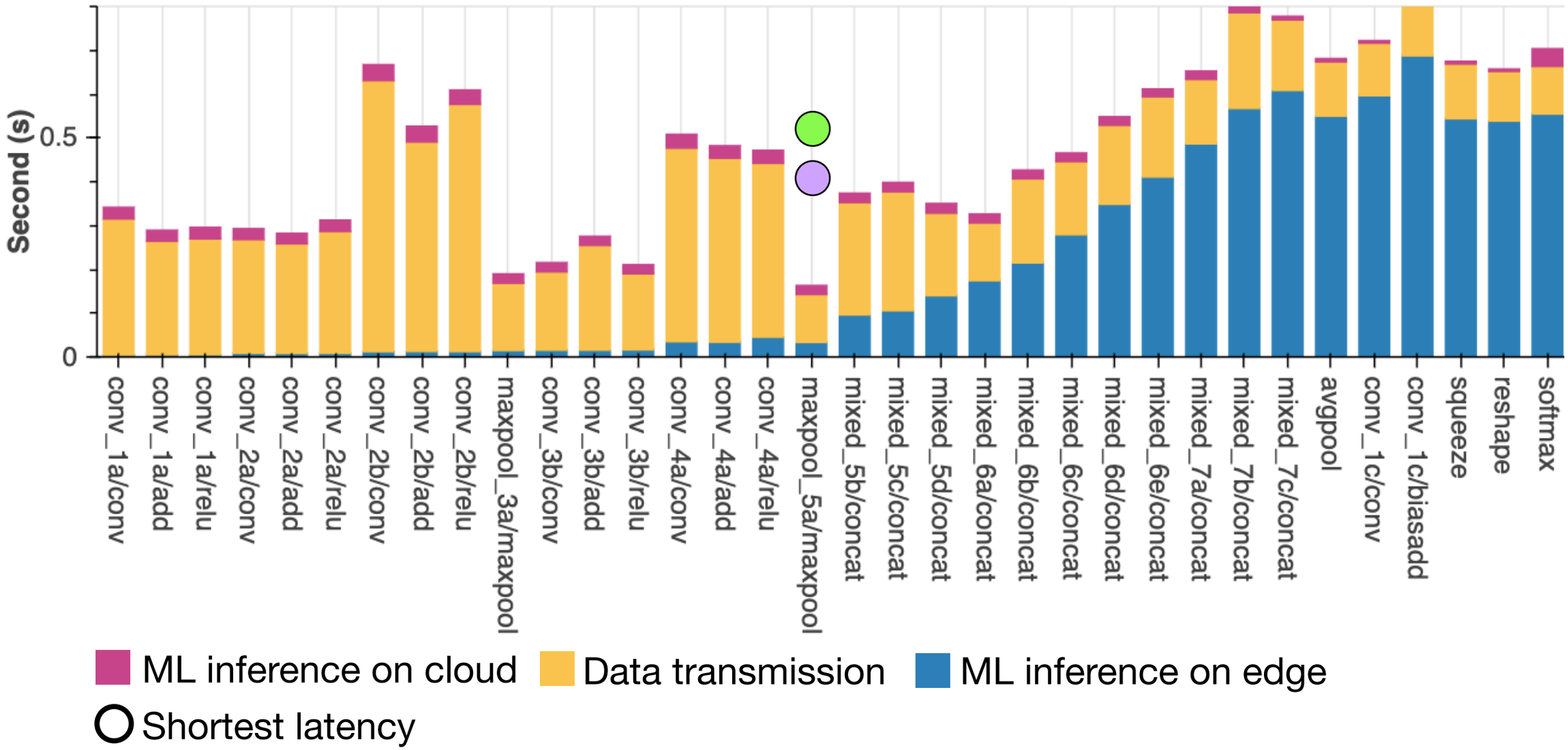
Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2



Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2

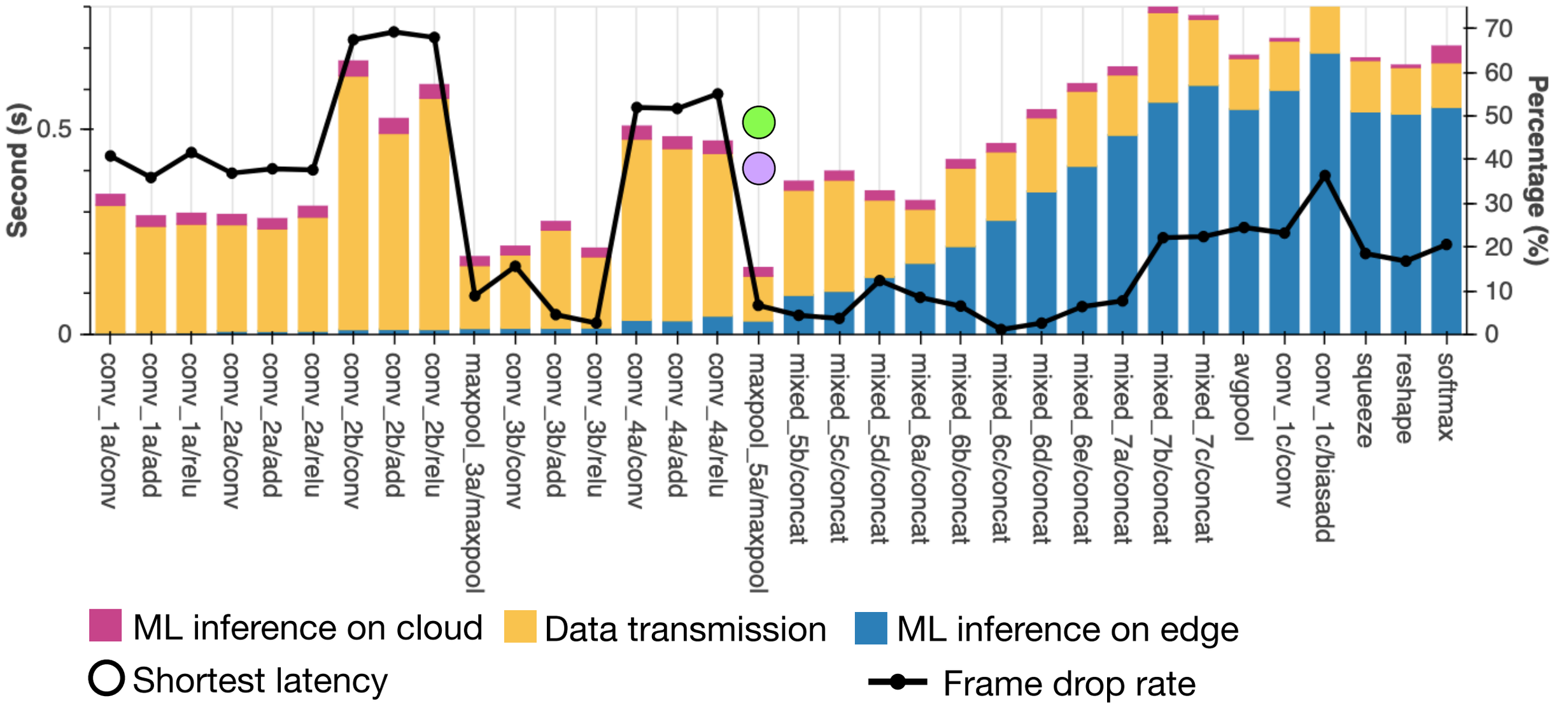


Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2

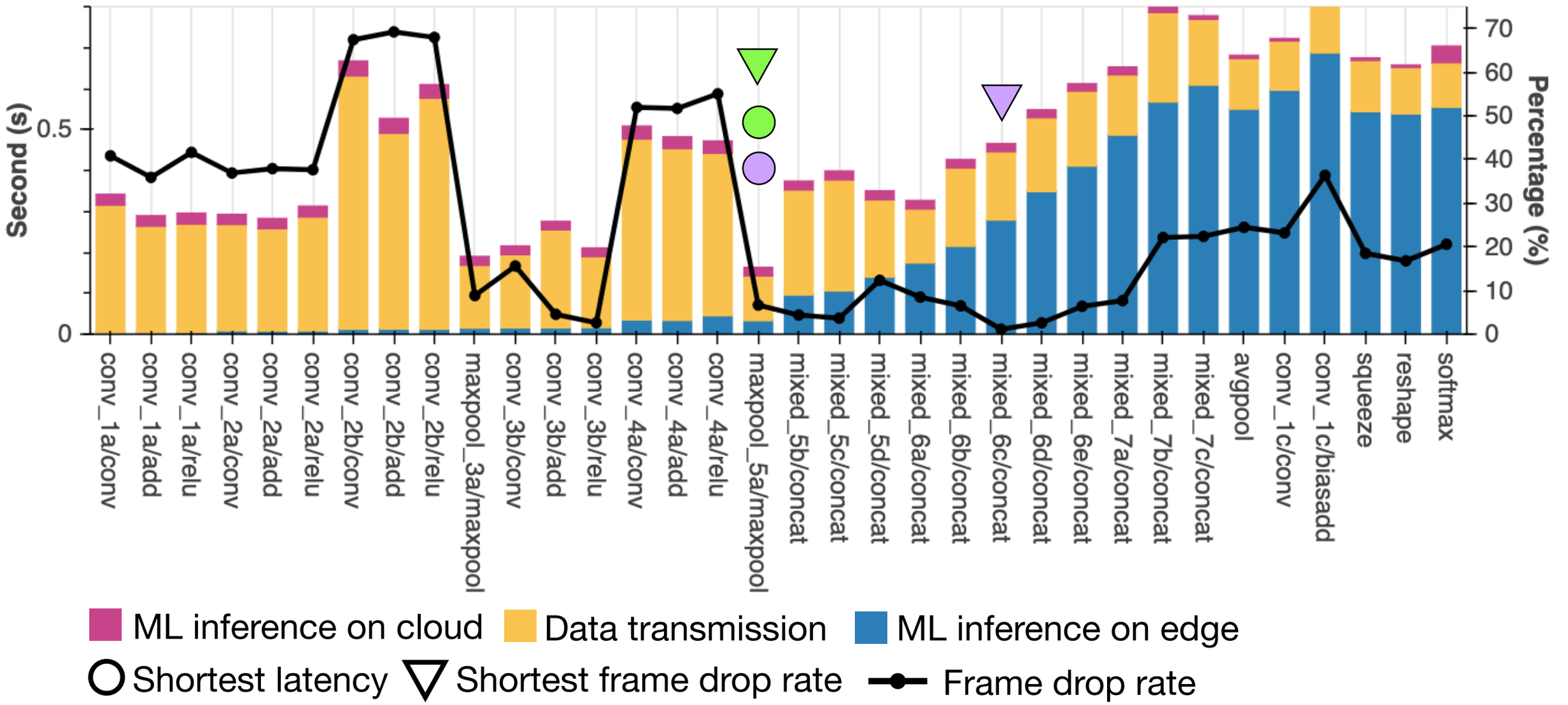


Hybrid method can find the same slicing deployment with much smaller problem space

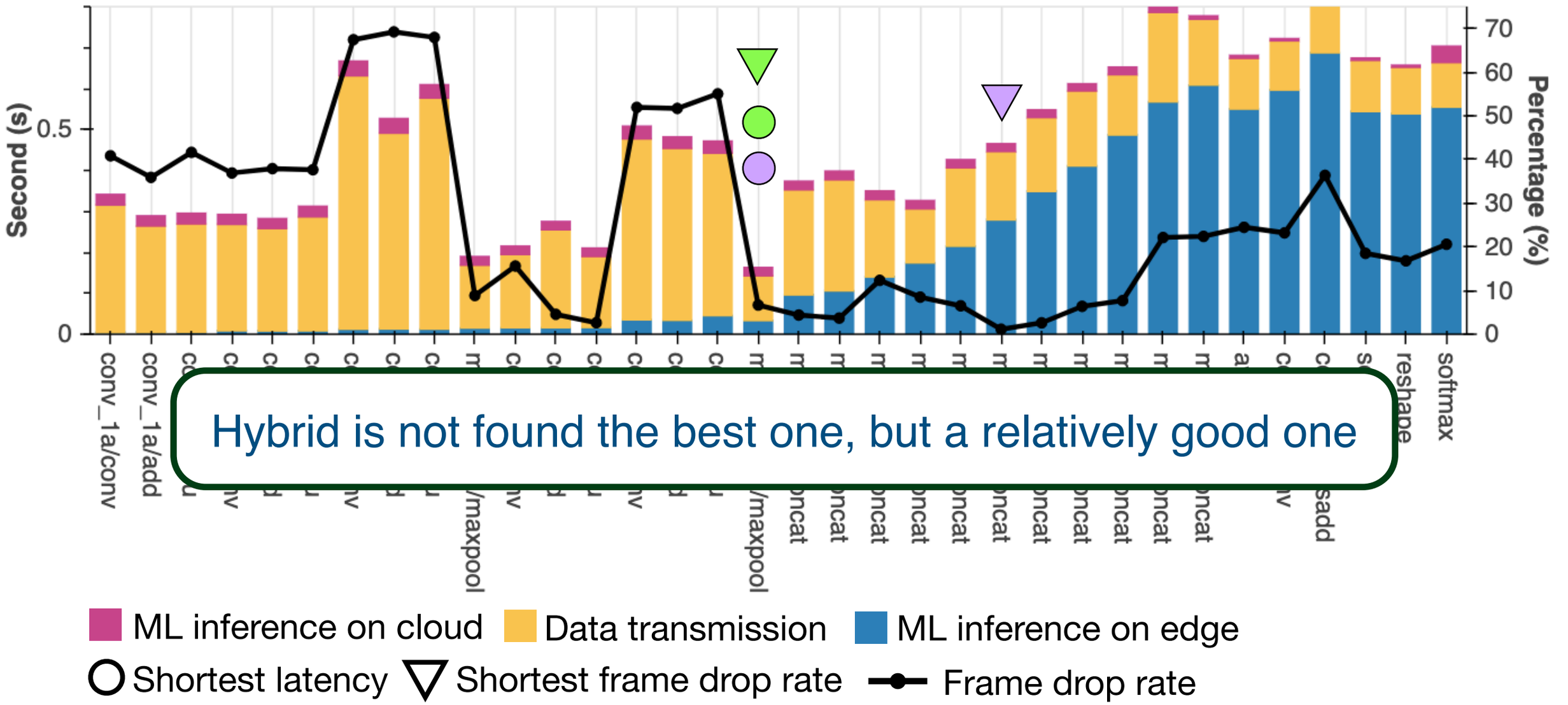
Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2



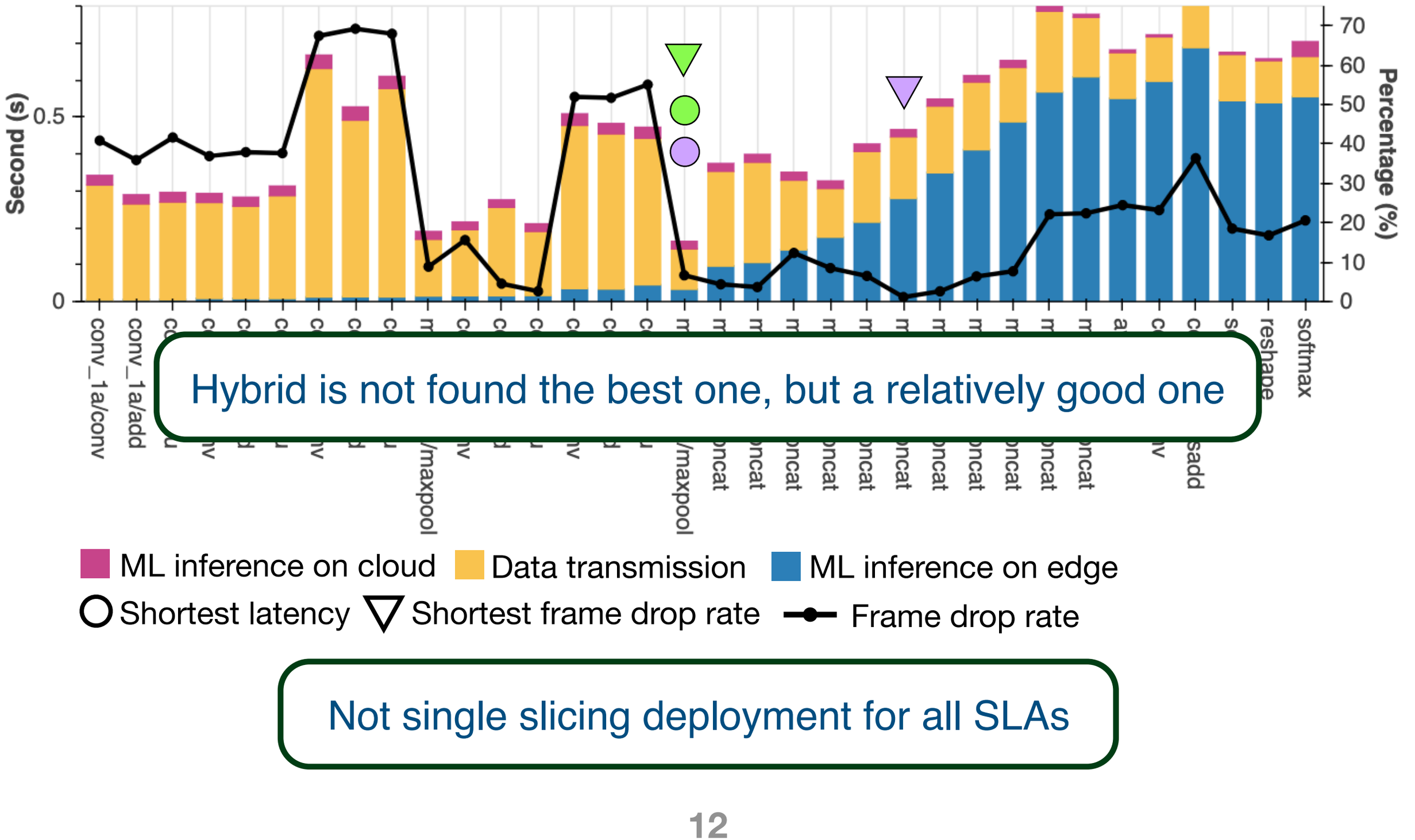
Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2



Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2



Model	# Operator	Method	
		Strongman	Hybrid
Inception V3	788	34	2



Not single slicing deployment for all SLAs

Couper contribution

- **Improve DNN inference on various metrics**:
  Achieved up to **90%** improvement on processing latency and
  100% improvement on processing quality.

- **Rapid to find solution**:
  Reduced **99%** problem space for searching best deployment.

- **Flexible to different DNN inference service**:
  Supported pluggable slicing algorithm and evaluating method.

- **Compatible with contemporary software stack**:
  Deployed with container orchestration, Kubernetes.

Thanks for your attention!

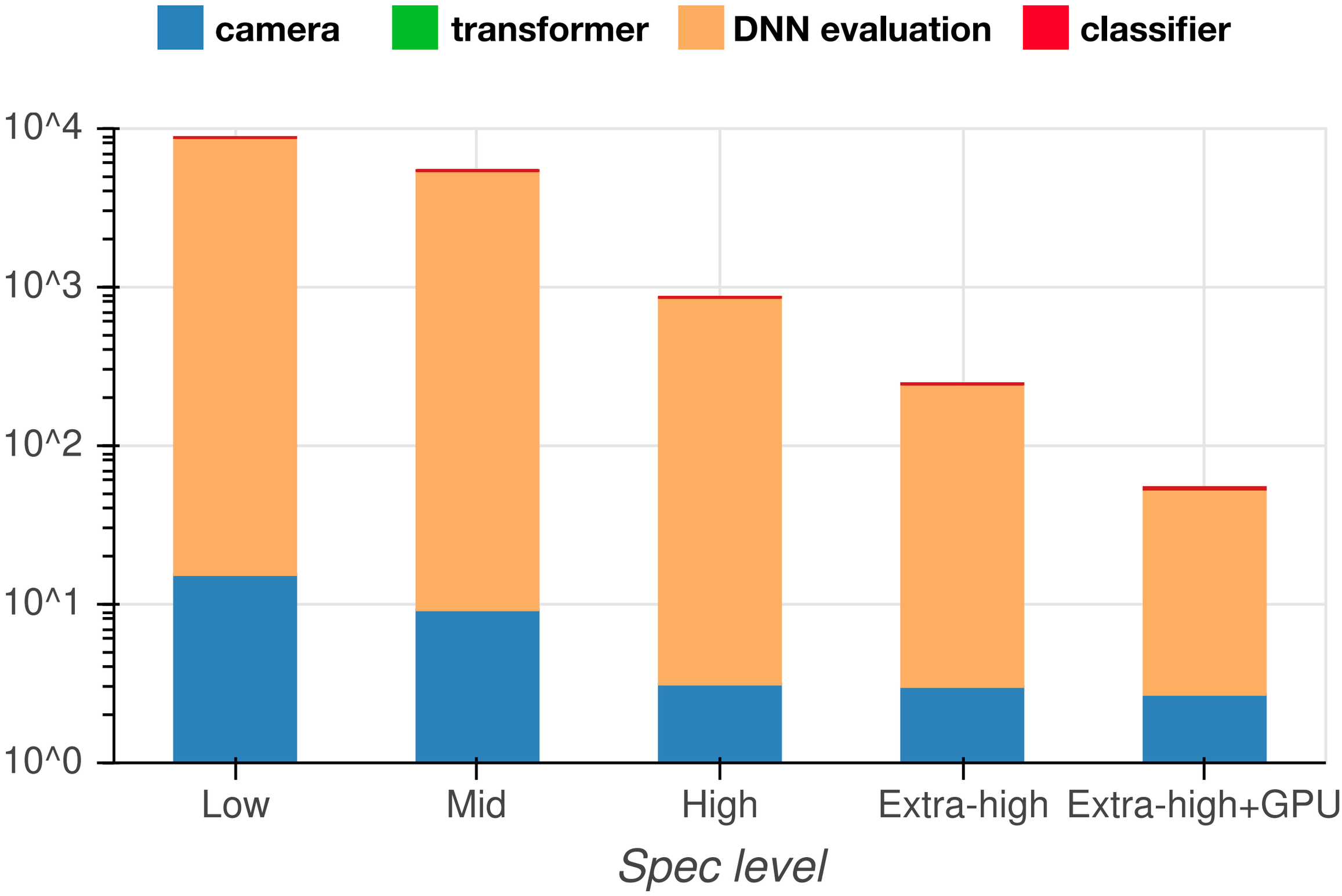
Hewlett Packard Enterprise

intel

vmware

ada
Applications Driving Architectures

Running PNASNet on different edge



Here comes Couper !

This is a multi-dimensional problem:

1. Heterogeneous computing resource between client, edge and cloud.
2. Various compute-intensive DNN models

=> slicing the DNN to fit the edge resource

Here comes Couper !

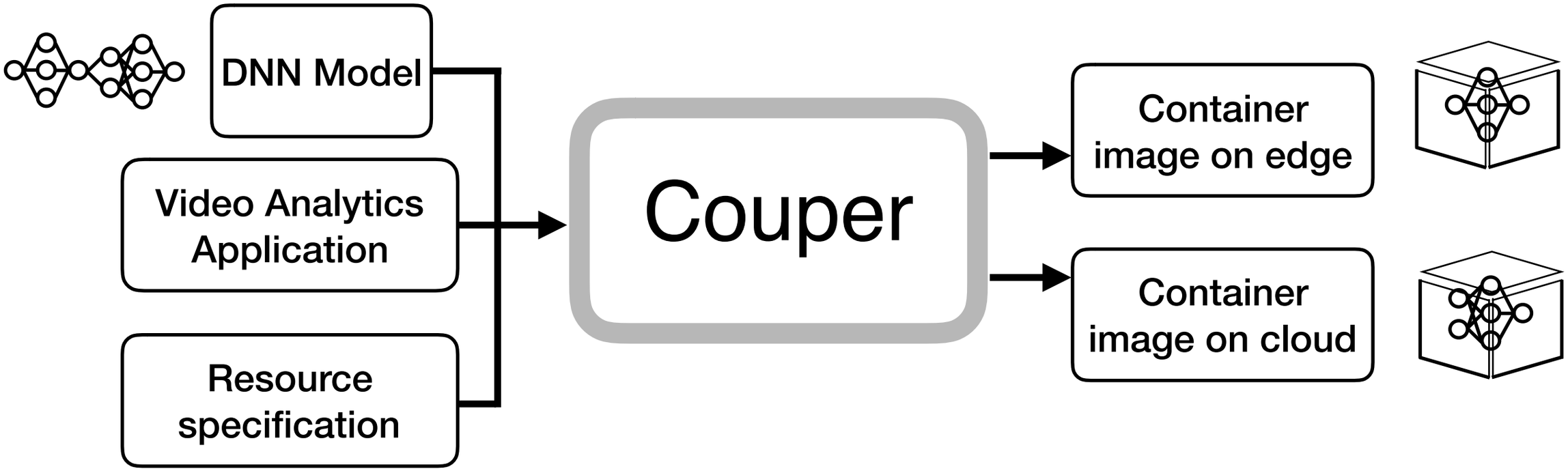
This is a multi-dimensional problem:

1. Heterogeneous computing resource between client, edge and cloud.
2. Various compute-intensive DNN models

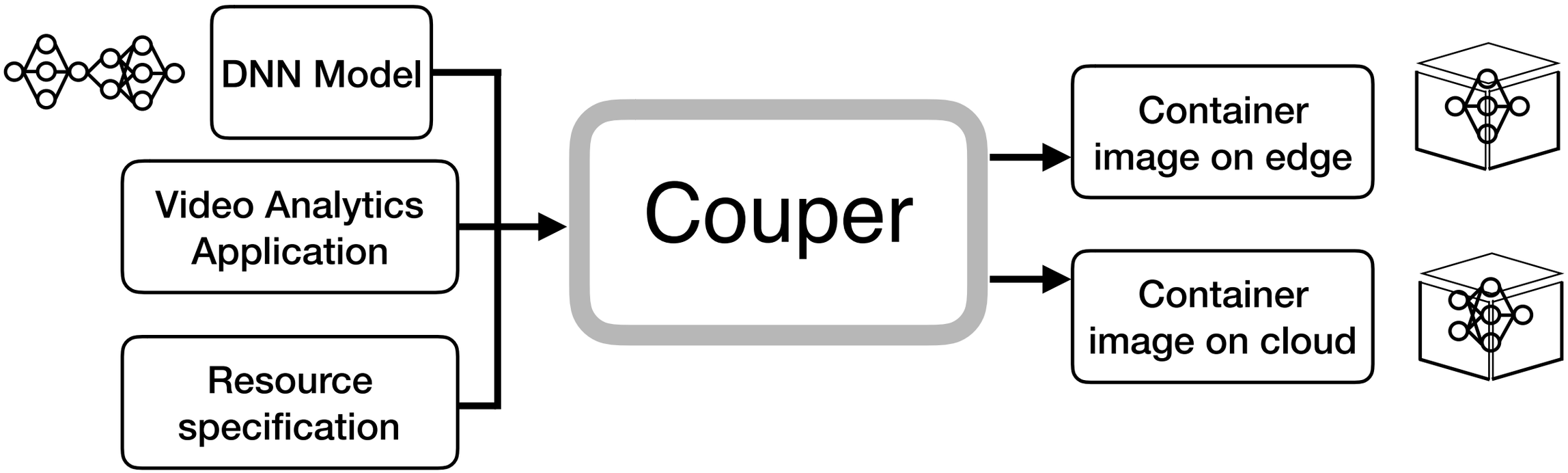
=> slicing the DNN to fit the edge resource

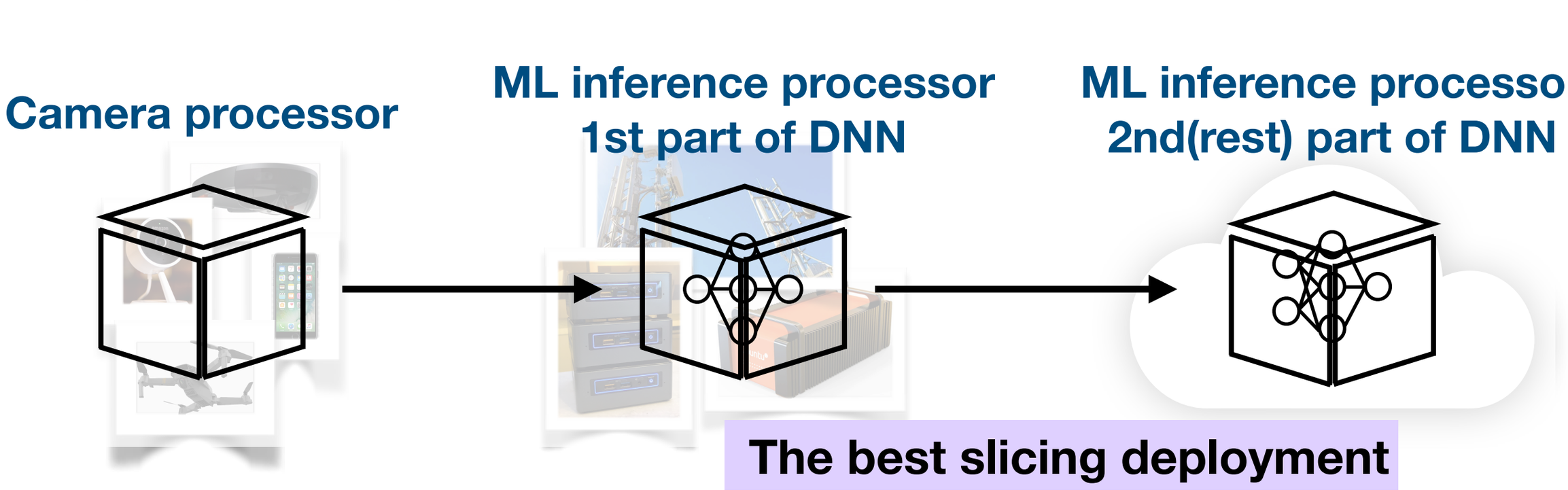
	Neurosurgeon (ASPLOS'17)	DDNN (ICDCS'17)	Edge-host partitioning of DNN (AVSS'18)	**Couper**
Edge involved?		✓	✓	✓
Generic slicing method?				✓
Verified by production DNN?	✓		✓	✓
Supporting different tenancies?				✓

Couper Introduction



Couper Introduction





Experiment

Goals:

- How Couper **improves performance**?
- How Couper **reduces problem space** and saves evaluation time?
- Why Couper **supports different evaluating methods**?

Hardware specification of experiments:

Device	CPU Freq (GHz)	CPU proc	RAM (GB)	GPU	RTT (ms)	
					client	cloud
Client device	2.0	2	1	N/A		
Low-end edge	2.0	4	16		1	65
Mid-end edge	3.1	8	32		15	50
High-end edge	3.1	16	64		25	42
Cloud server	3.1	48	96	2 Nvidia P100		

Experiment

Goals:

- How Couper **improves performance**?
- How Couper **reduces problem space** and saves evaluation time?
- Why Couper **supports different evaluating methods**?

Hardware specification of experiments:

Device	CPU Freq (GHz)	CPU proc	RAM (GB)	GPU	RTT (ms)	
					client	cloud
Client device	2.0	2	1	N/A		
Low-end edge	2.0	4	16		1	65
Mid-end edge	3.1	8	32		15	50
High-end edge	3.1	16	64		25	42
Cloud server	3.1	48	96	2 Nvidia P100		

More powerful edge is further from client

Experiment

The original layers of DNN and the # evaluation candidates

Model	# Layer	Method		
		Strongman	Comm-slim	Hybrid
VGG 16	54	52	20	1
MobileNet V2 1.4	158	155	132	3
ResNet V2 50	205	34	15	1
Inception V3	788	34	15	2
Inception ResNet V2	871	106	28	3
NASNet 331	1265	7	3	1
PNASNet 331	939	7	3	1

Experiment

The original layers of DNN and the # evaluation candidates

Model	# Layer	Method		
		Strongman	Comm-slim	Hybrid
VGG 16	54	52	20	1
MobileNet V2 1.4	158	155	132	3
ResNet V2 50	205	34	15	1
Inception V3	788	34	15	2
Inception ResNet V2	871	106	28	3
NASNet 331	1265	7	3	1
PNASNet 331	939	7	3	1

Up to 98% evaluation time reduction

Hugely reduce problem space(split point candidates) by methods

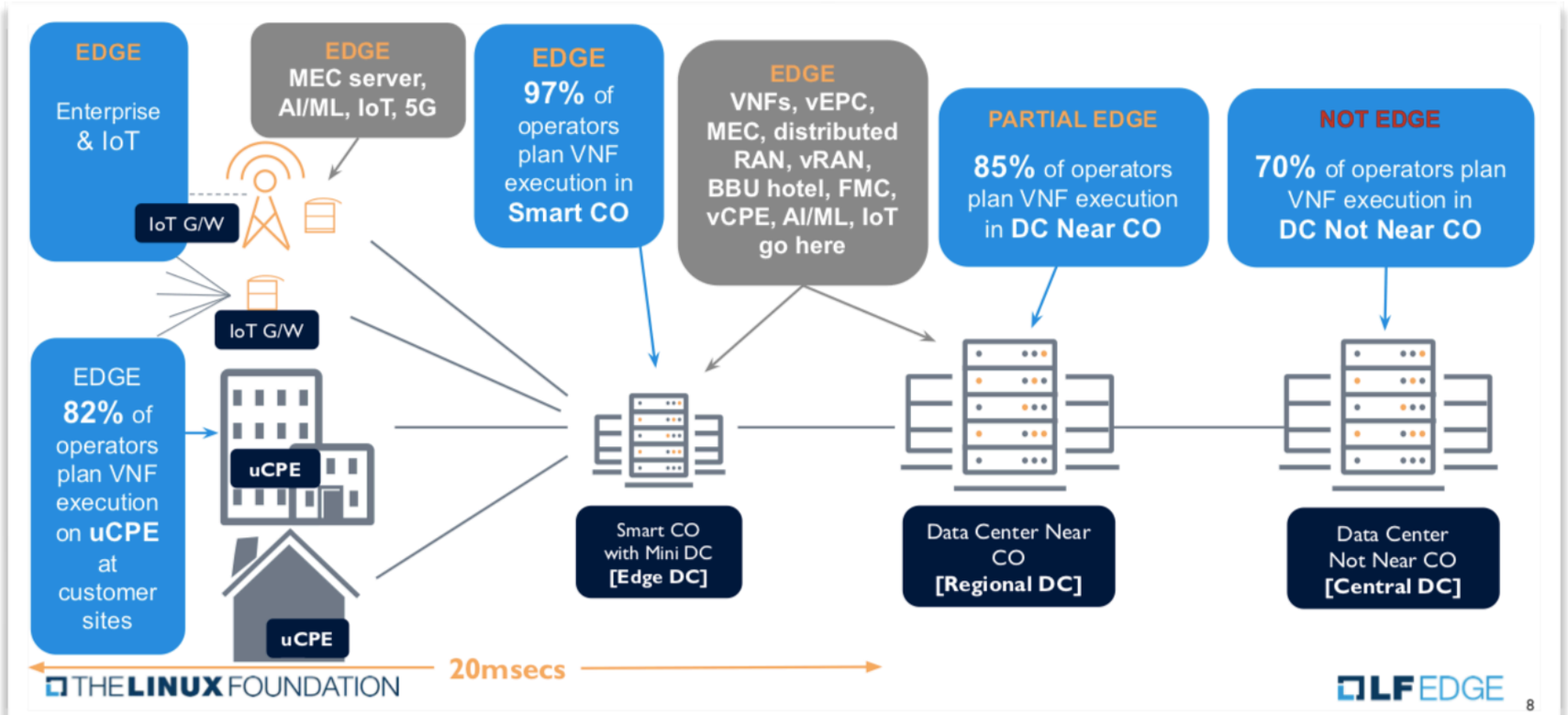
Next Step

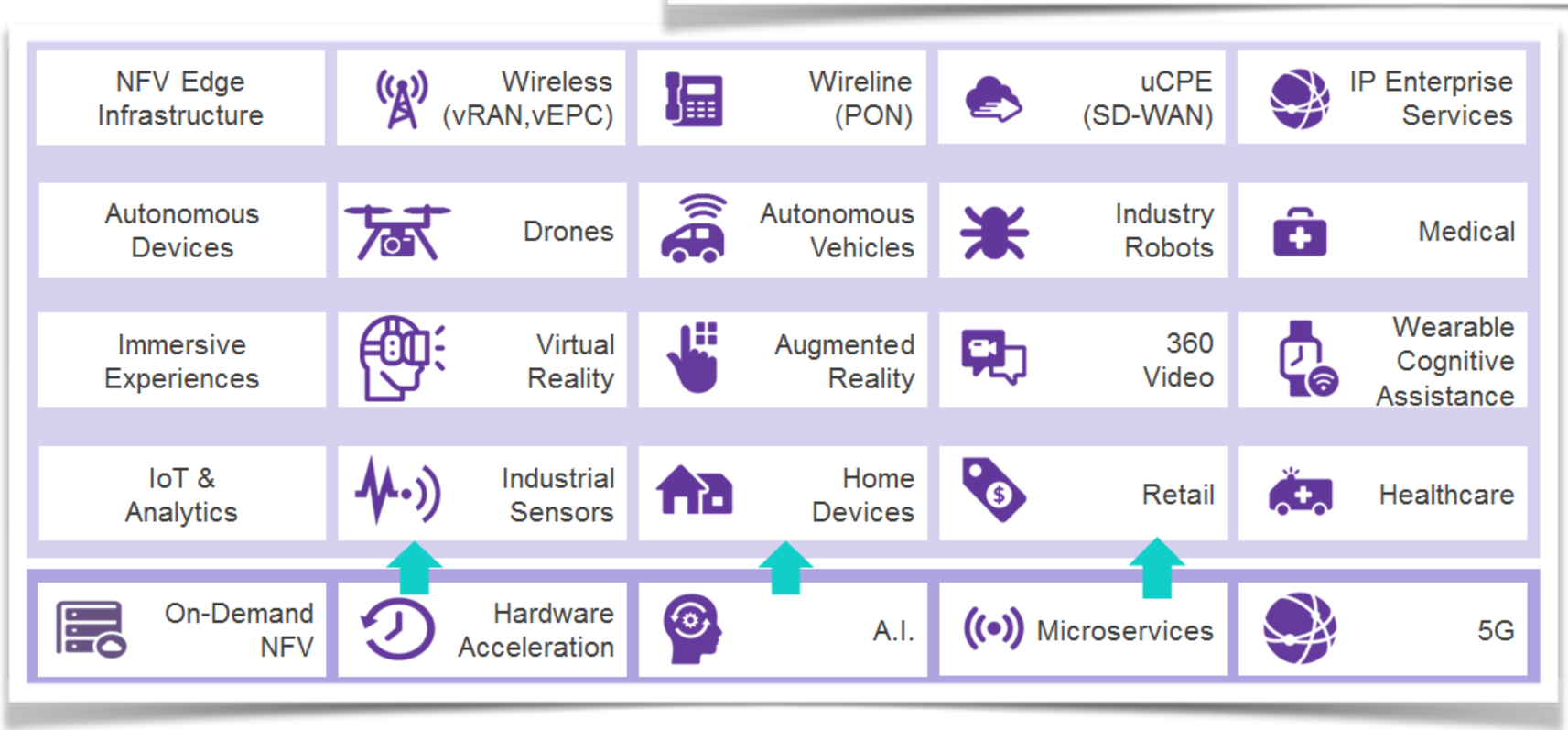
- **Couper Enhancement**:
  Working with different DNN model, application, and framework (i.e. Yolov3 with object detection)

- **Collaborate with edge software stack**:
  Evaluating 5G environment, edge infrastructure (i.e. Akraino), and supporting software (i.e. NFV techniques)

- **Multi-tenancy with different workloads**:
  Evaluating on the compute and network interference/overhead while sharing resource with other services

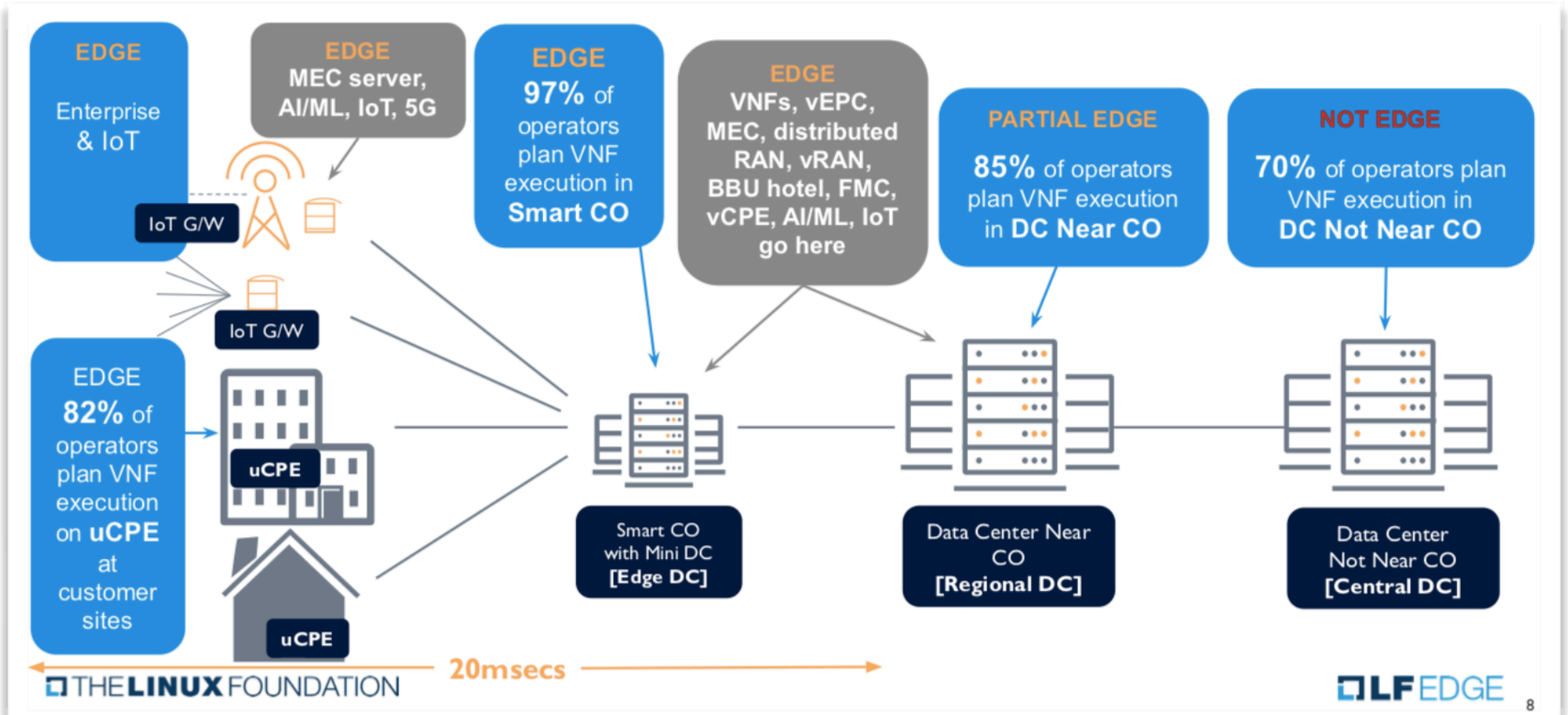
Edge resources are diverse and target to support multi-tenancy (backup page)

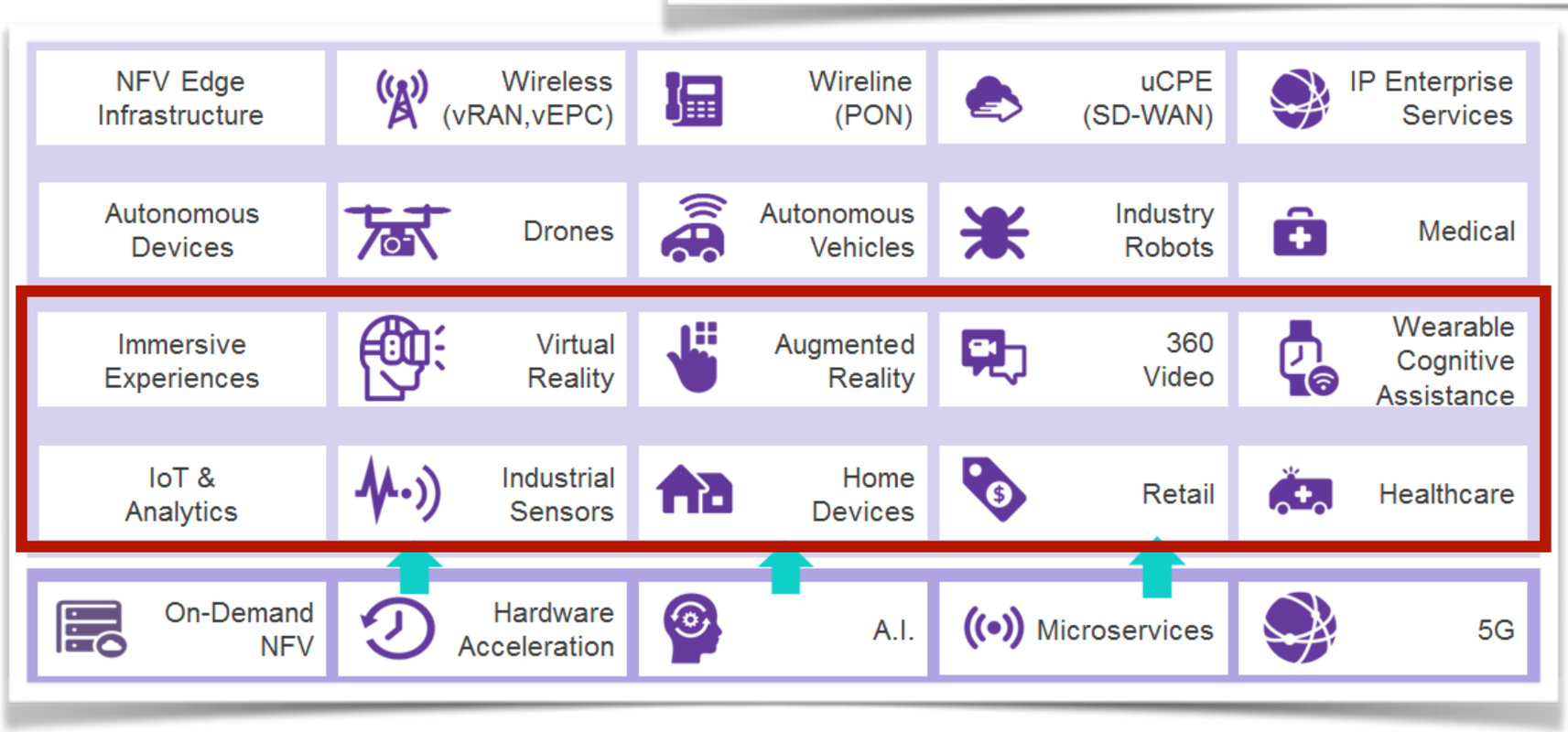




- Linux Foundation Edge, Akraino — emerging technology and edge coverage

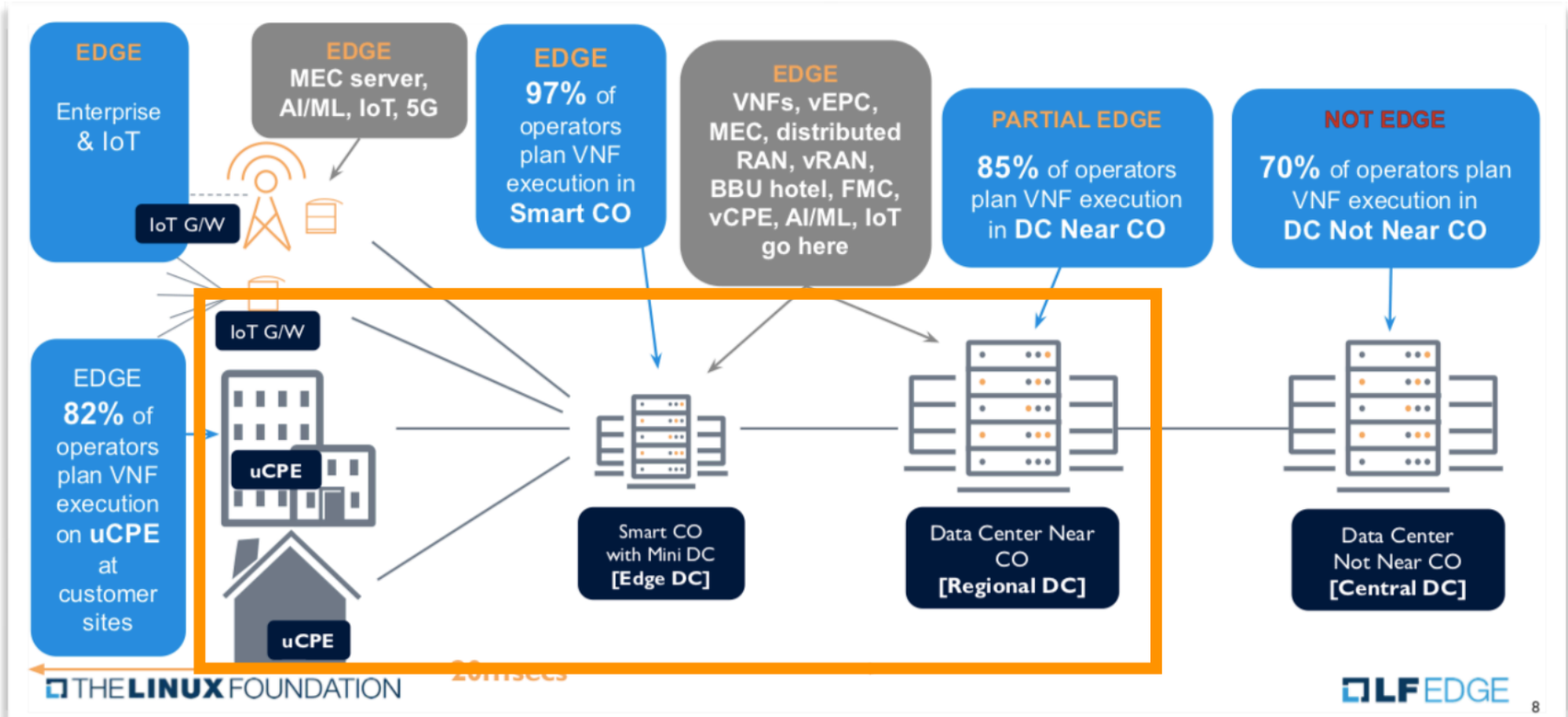
Edge resources are diverse and target to support multi-tenancy (backup page)

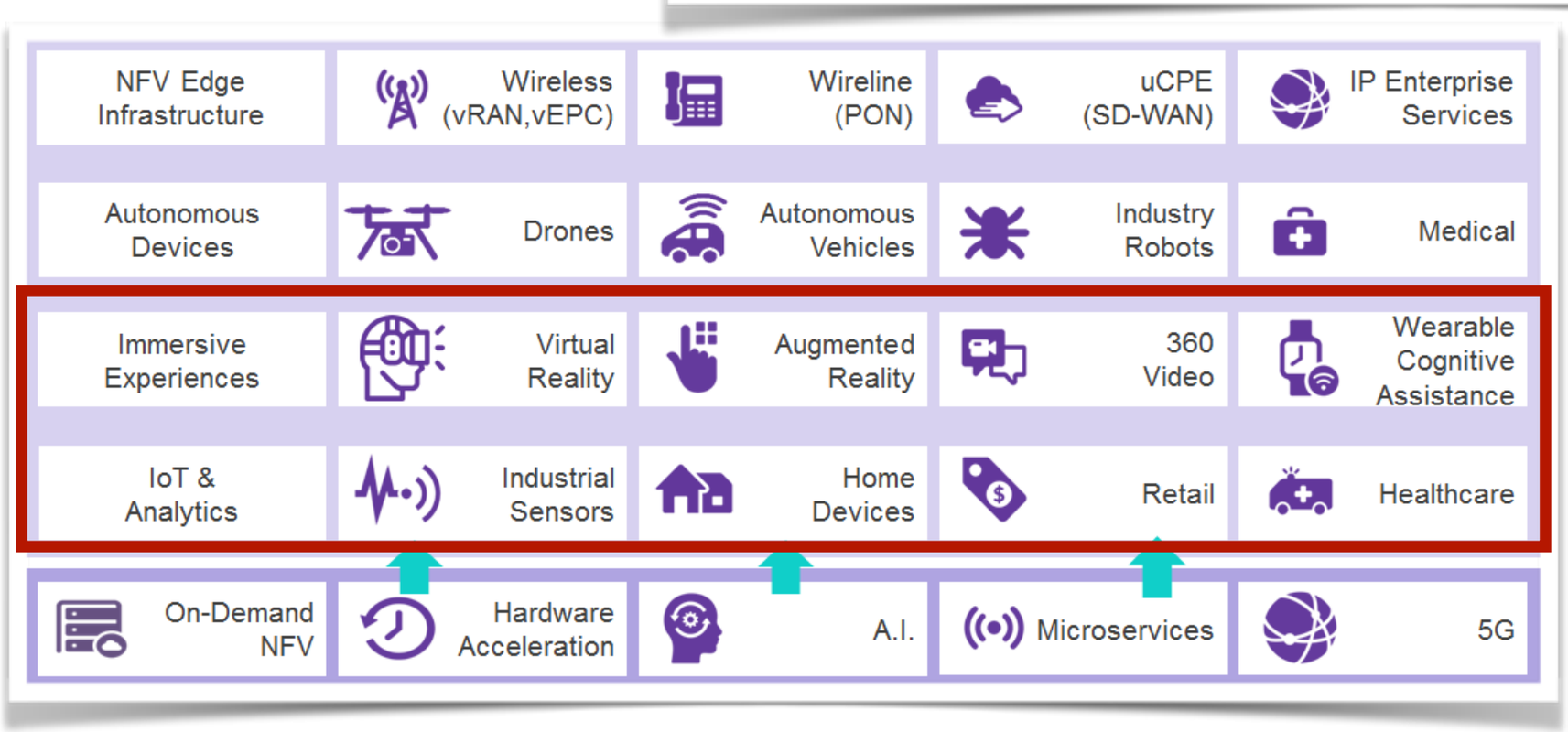




- Linux Foundation Edge, Akraino — emerging technology and edge coverage

Edge resources are diverse and target to support multi-tenancy (backup page)

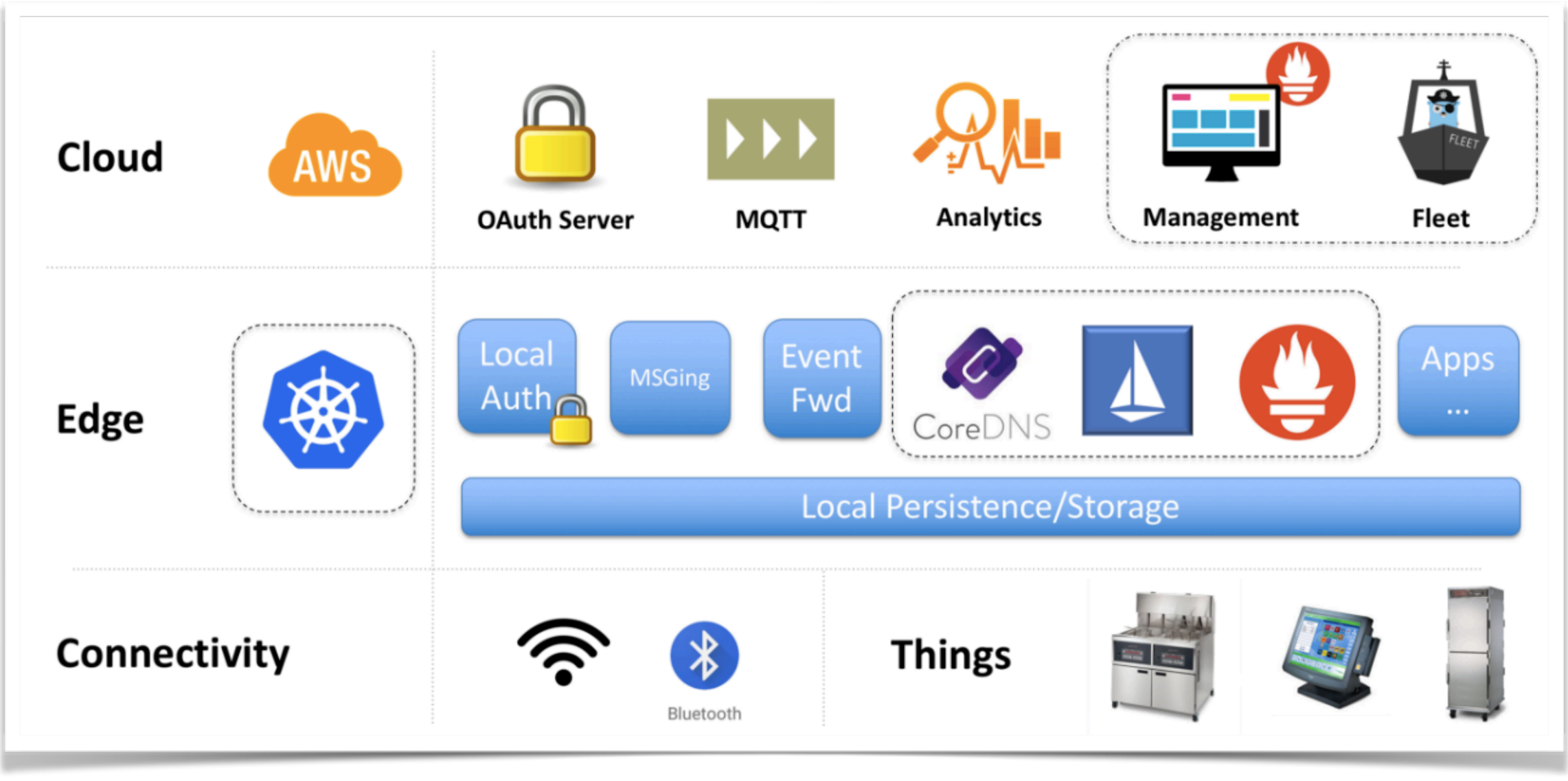




- Linux Foundation Edge, Akraino — emerging technology and edge coverage

Edge resources are diverse and target to support multi-tenancy

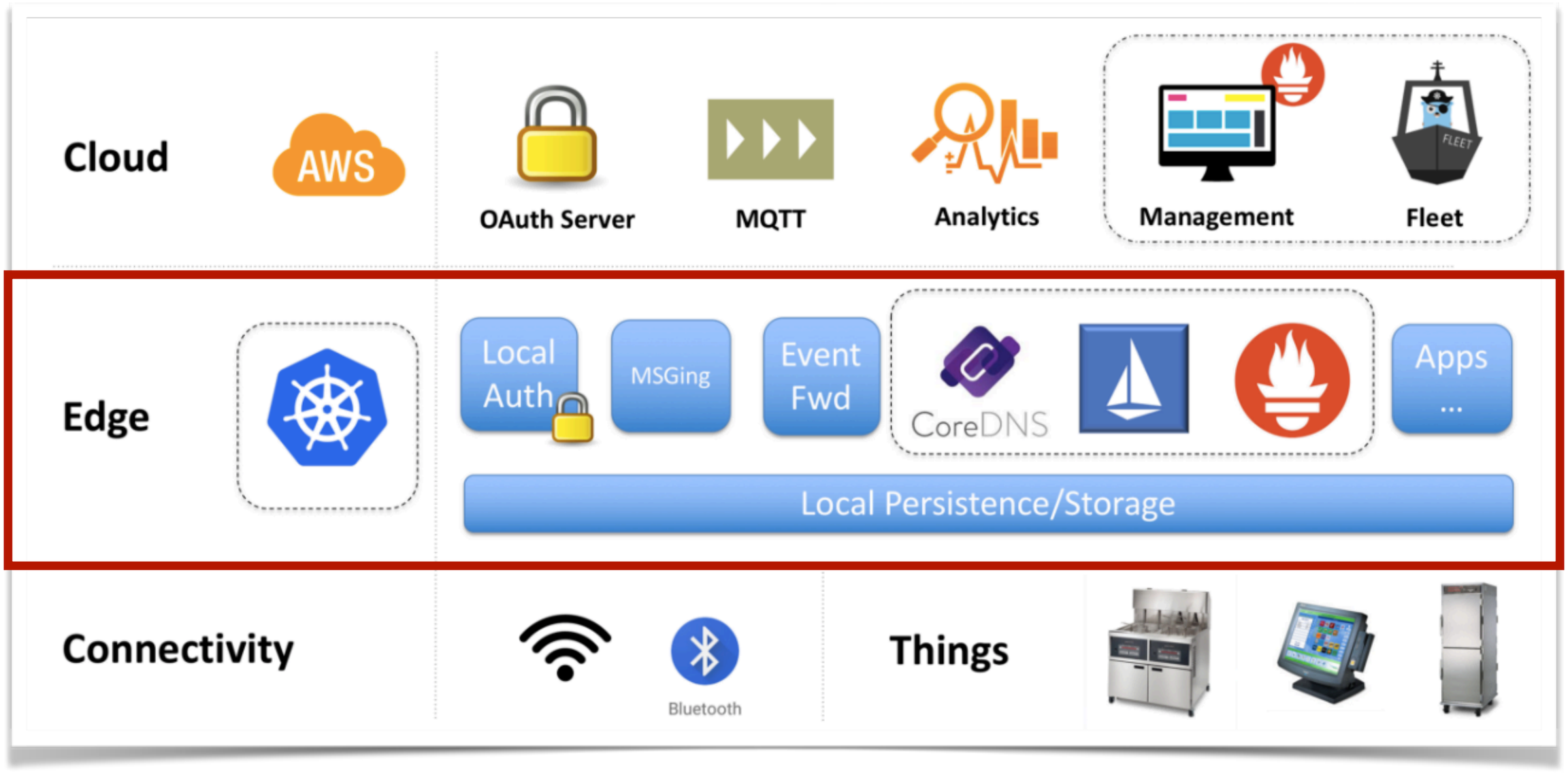
Even in specific edge device owned by certain company, need to support multiple services



❖ Chick-fil-A, Edge computing architecture overview

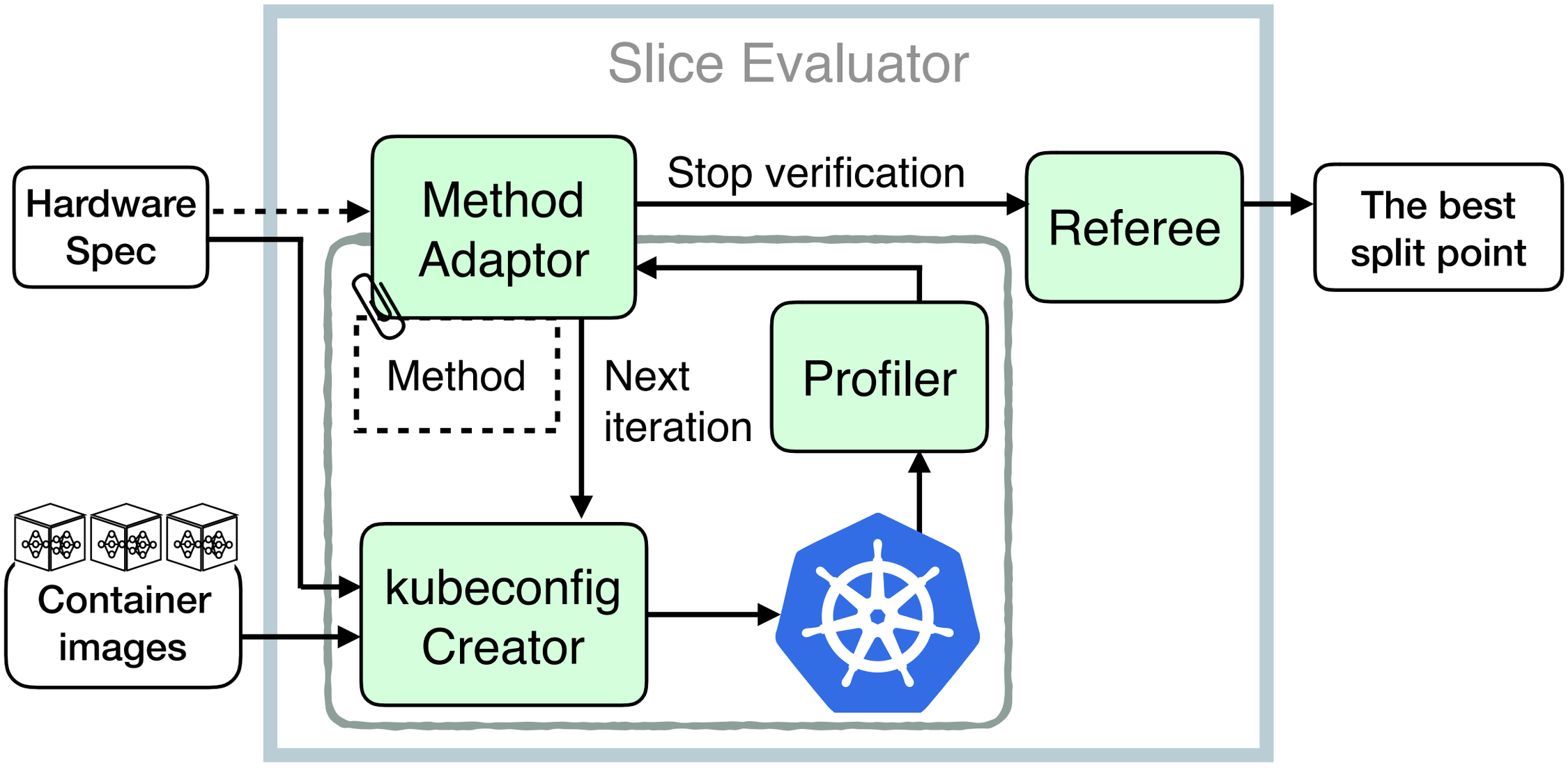
Edge resources are diverse and target to support multi-tenancy

Even in specific edge device owned by certain company, need to support multiple services



❖ Chick-fil-A, Edge computing architecture overview

Couper Introduction



Experiment

Goals:

- How Couper **improves performance**?
- How Couper **reduces problem space** and saves evaluation time?
- Why Couper **supports different evaluating methods**?

Hardware specification of experiments:

Device	CPU Freq (GHz)	CPU proc	RAM (GB)	GPU	RTT (ms)	
					client	cloud
Client device	2.0	2	1	N/A		
Low-end edge	2.0	4	16		1	65
Mid-end edge	3.1	8	32		15	50
High-end edge	3.1	16	64		25	42
Super-high-end edge	3.1	16	64	1 Nvidia P100	25	42
Cloud server	3.1	48	96	2 Nvidia P100		

Experiment

Goals:

- How Couper **improves performance**?
- How Couper **reduces problem space** and saves evaluation time?
- Why Couper **supports different evaluating methods**?

Hardware specification of experiments:

Device	CPU Freq (GHz)	CPU proc	RAM (GB)	GPU	RTT (ms)	
					client	cloud
Client device	2.0	2	1	N/A		
Low-end edge	2.0	4	16		1	65
Mid-end edge	3.1	8	32		15	50
High-end edge	3.1	16	64		25	42
Super-high-end edge	3.1	16	64	1 Nvidia P100	25	42
Cloud server	3.1	48	96	2 Nvidia P100		

More powerful edge is further from client

(backup page)

Evaluation

Real evaluation time in minutes across models and edge devices, the hybrid method comes out decision more faster than strongman

Model	Inception V3	Inception ResNet V2	PNASNet 331
The evaluation time of Strongman	> 30	≈ 120	≈ 10
Low-end edge	1	1	1
Mid-end edge	2	3	1
High-end edge	10	16	1