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

ACM/IEEE Symposium on Edge Computing（SEC'19）

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**

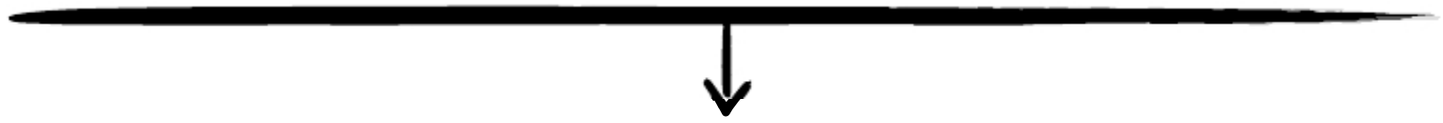
How does video analytics application work with edge?

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	VGG16	MobileNet V21.4	ResNetV2 50	Inception V3	Inception ResNetV2	NASNet 331	PNASNet 331
Released Time	2014Sep	2018Jan	2016Jul	2016Jul	2016Aug	2018Apr	2018Jul
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

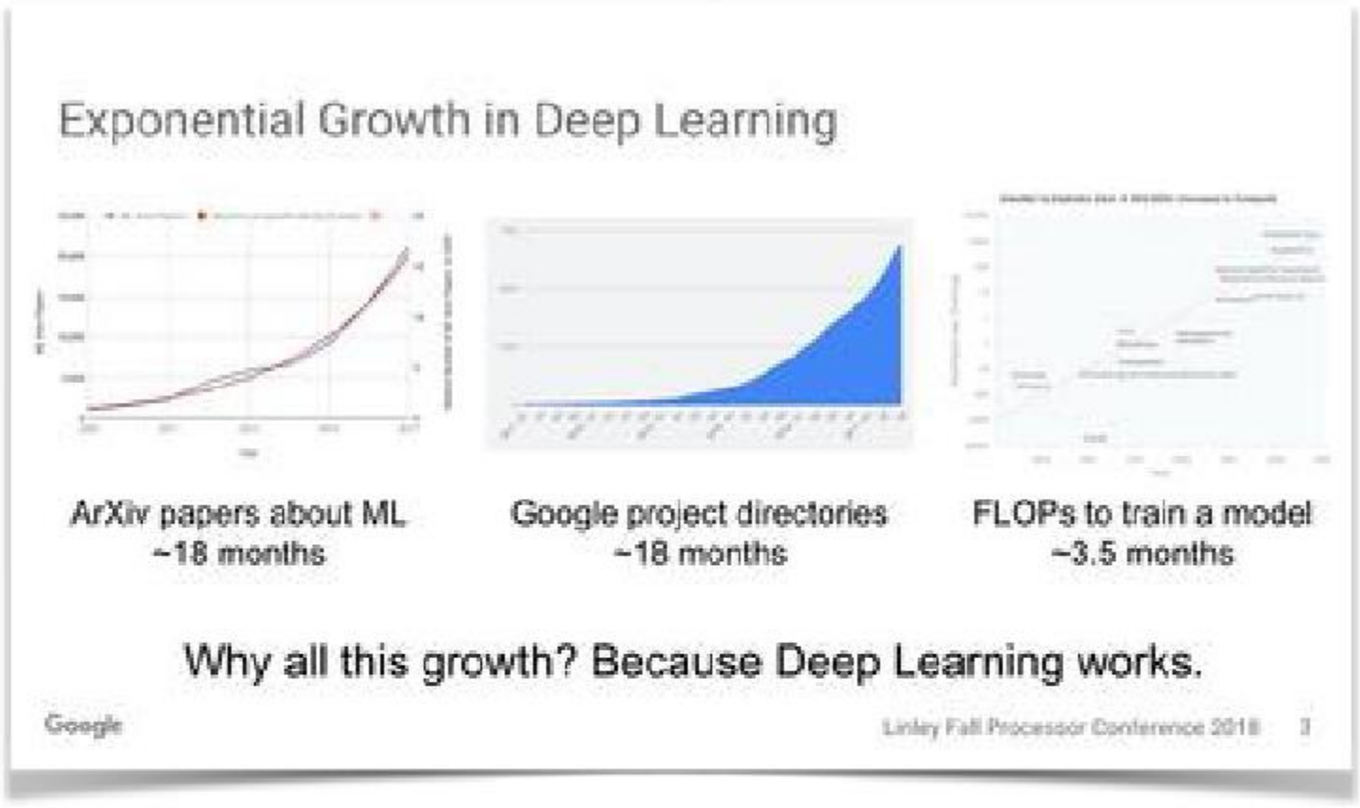
How does video analytics application work with edge?

Deep neural network (DNN)

- Google, Cliff Young
  (Linley Fall 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

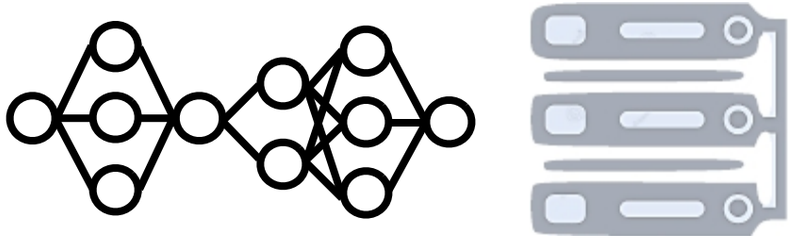
Bringing out edge's benefit is not easy

Bringing out edge's benefit is not easy

If edge cannot run whole DNN:

Bringing out edge's benefit is not easy

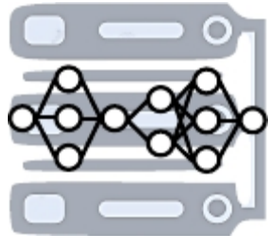
If edge cannot run whole DNN:



Optimize DNN for edge

Bringing out edge's benefit is not easy

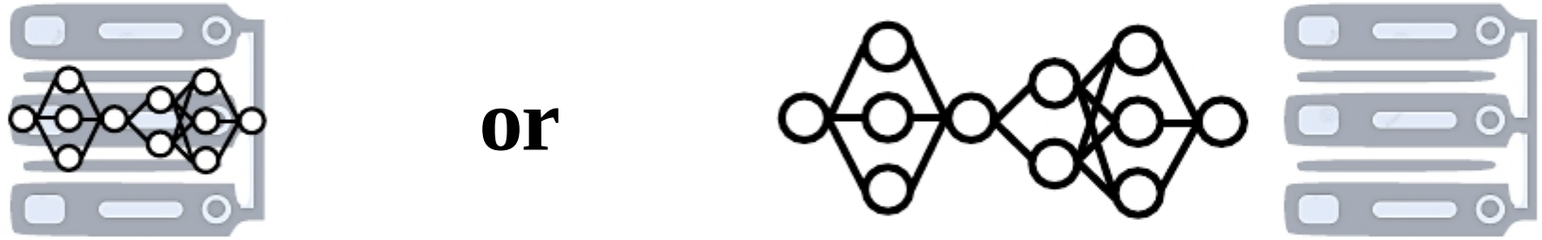
If edge cannot run whole DNN:



Optimize DNN for edge

Bringing out edge's benefit is not easy

If edge cannot run whole DNN:



Optimize DNN for edge **or** **Bring specific edge for DNN**

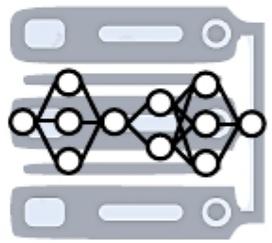
Bringing out edge's benefit is not easy

?

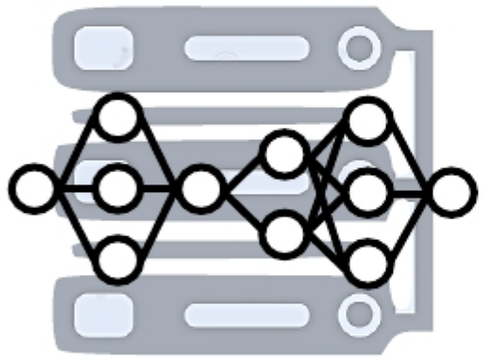
DNN

Client -> Edge -> Cloud

If edge cannot run whole DNN:



or



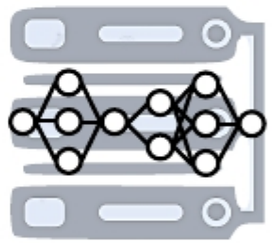
Optimize DNN for edge

Bring specific edge for DNN

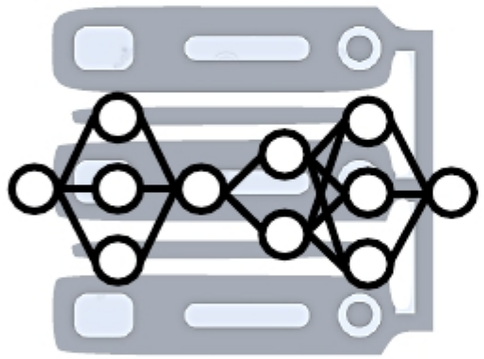
Bringing out edge's benefit is not easy

DNN ? Client -> Edge -> Cloud

If edge cannot run whole DNN:



or



Optimize DNN for edge

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**

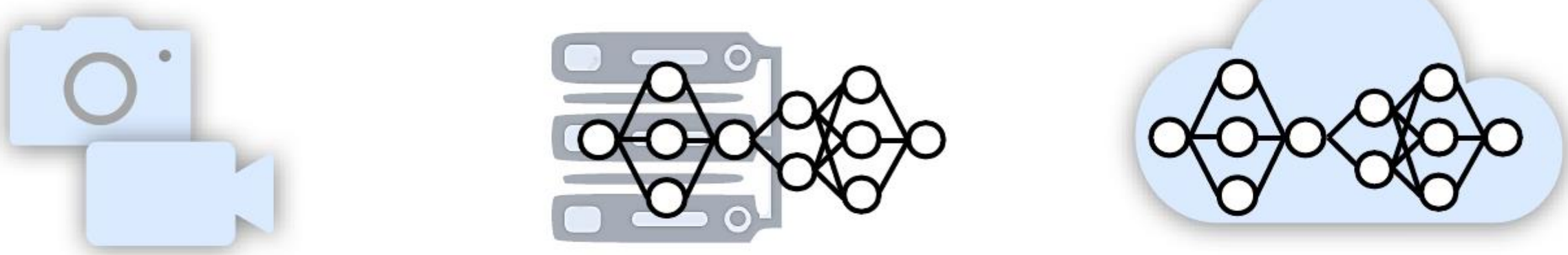
Given **a DNN** and an **edge**,

How can we deploy the model with good performance?

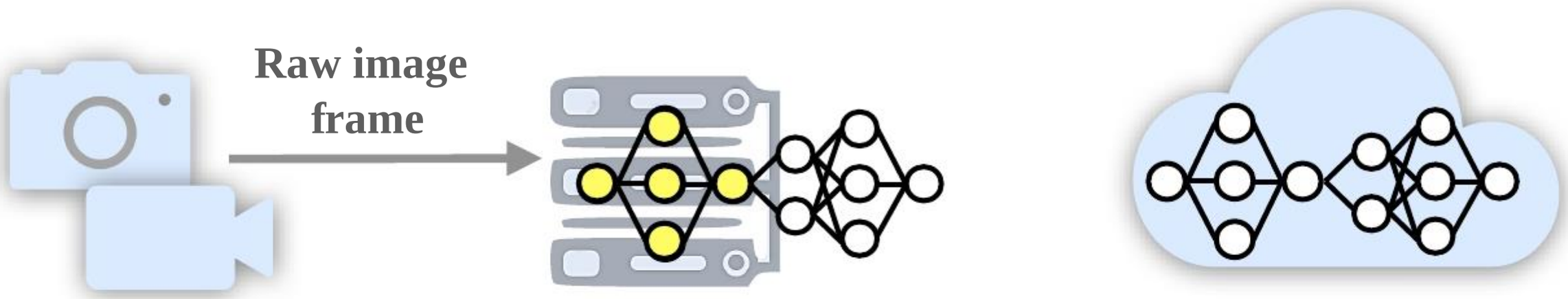
Couper: a general edge system

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

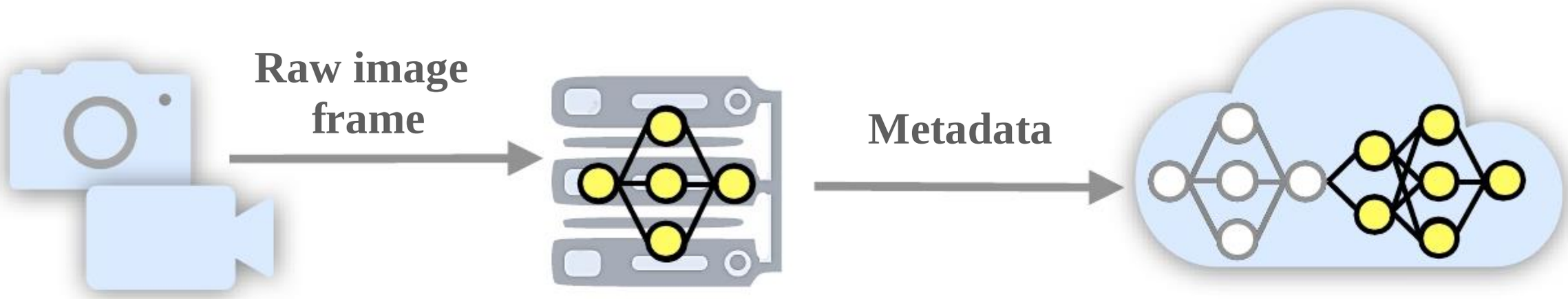
Share load across edge and cloud by DNN partitioning



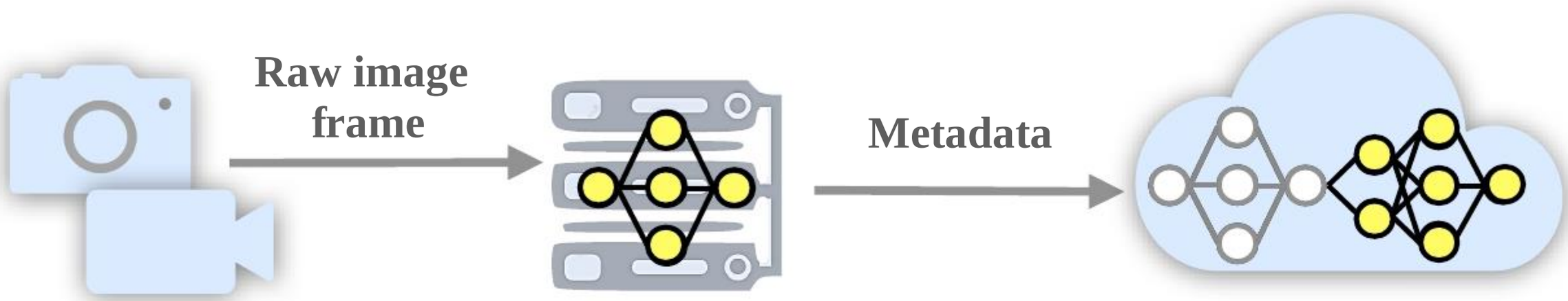
Share load across edge and cloud by DNN partitioning



Share load across edge and cloud by DNN partitioning

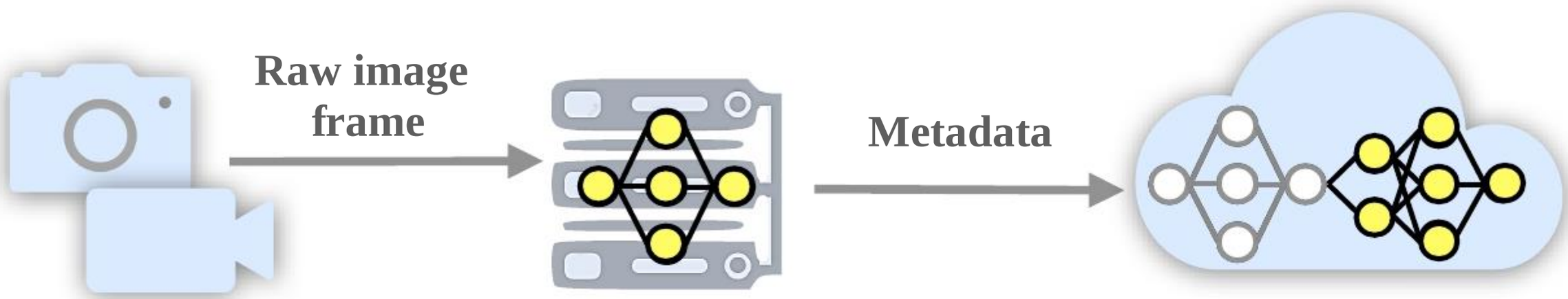


Share load across edge and cloud by DNN partitioning



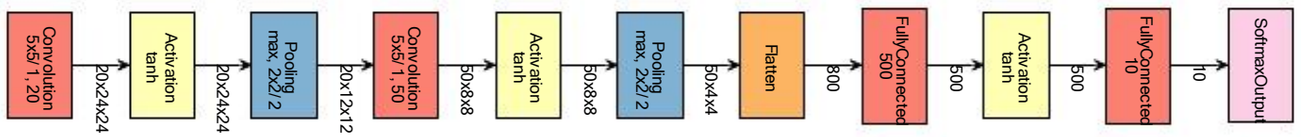
How do we decide the slicing point?

Share load across edge and cloud by DNN partitioning



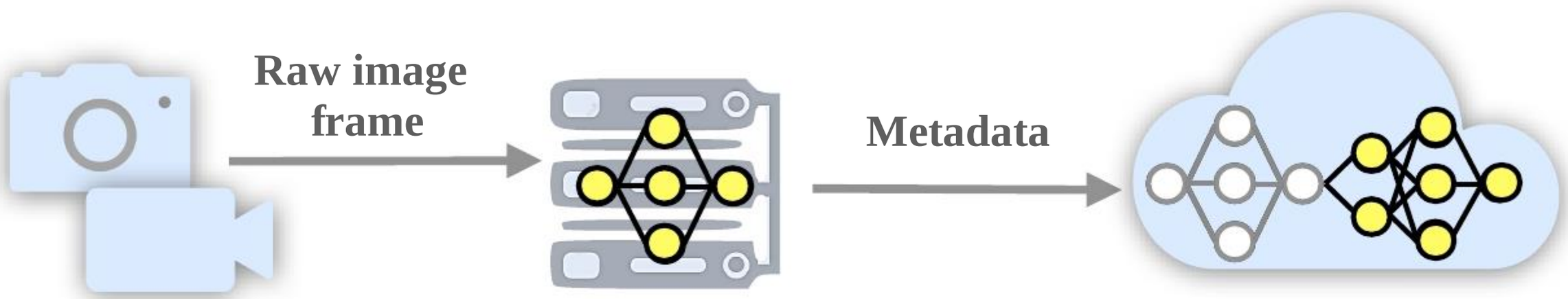
How do we decide the slicing point?

LeNet (1998)



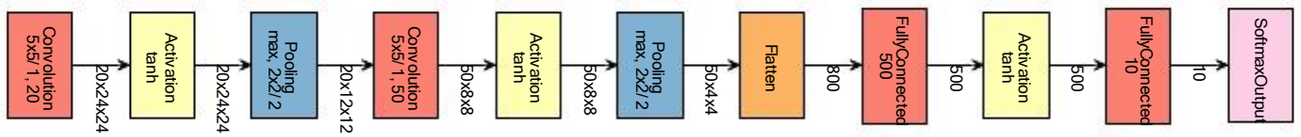
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning

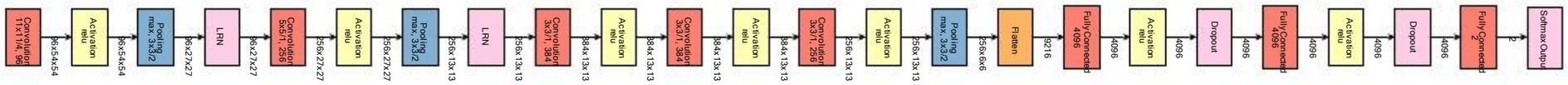


How do we decide the slicing point?

LeNet (1998)

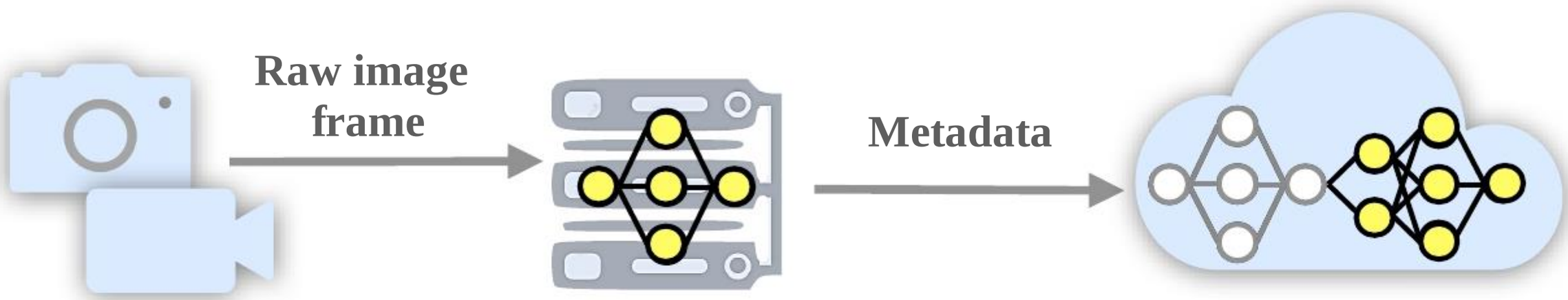


AlexNet (2012)



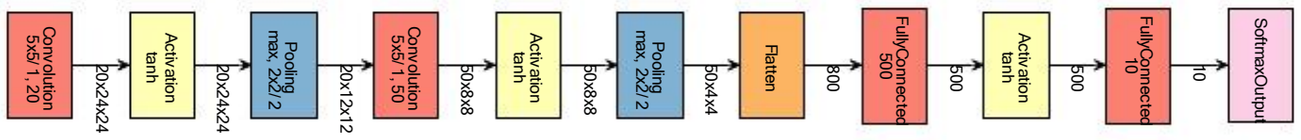
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning

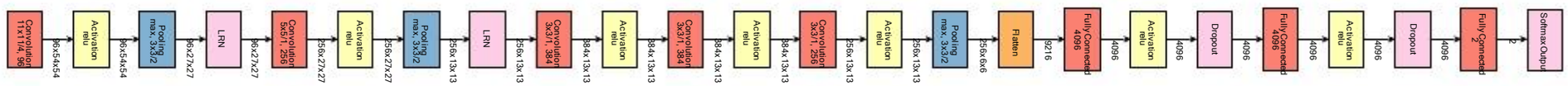


How do we decide the slicing point?

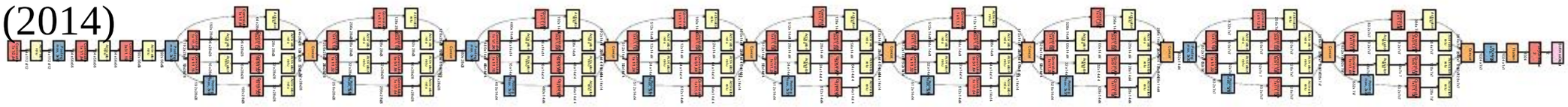
LeNet (1998)



AlexNet (2012)

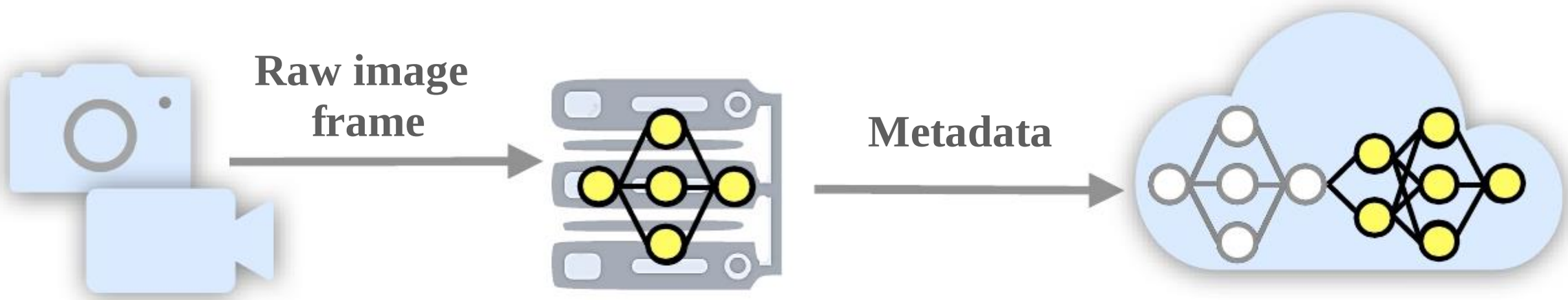


GoogLeNet (2014)



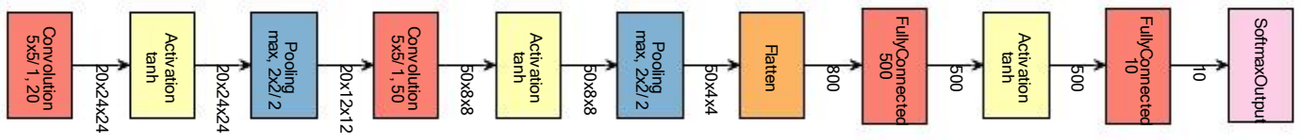
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning

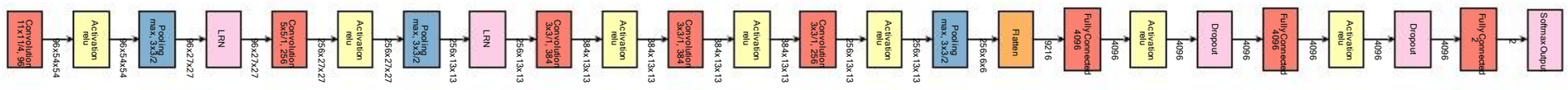


How do we decide the slicing point?

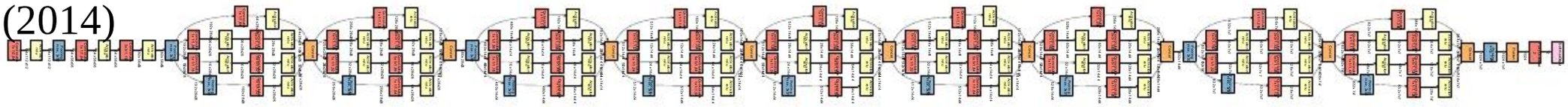
LeNet (1998)



AlexNet (2012)



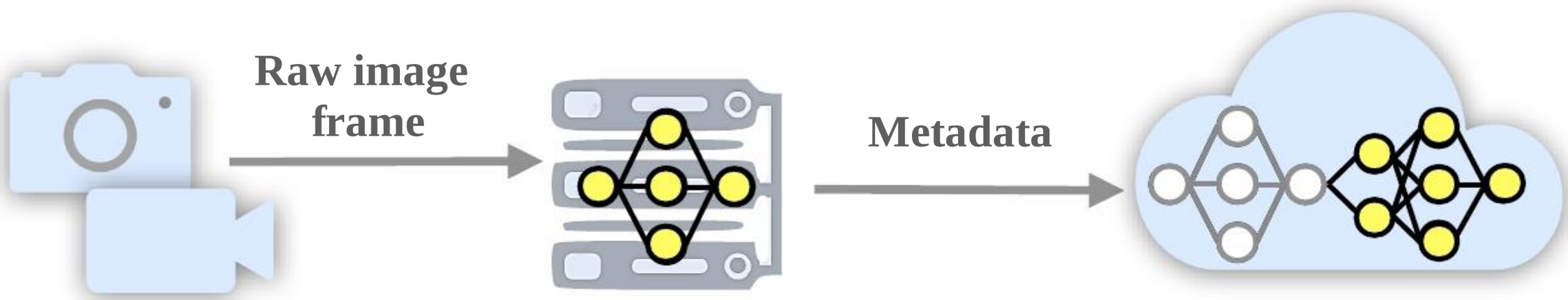
GoogLeNet (2014)



Inception V3 (2015)

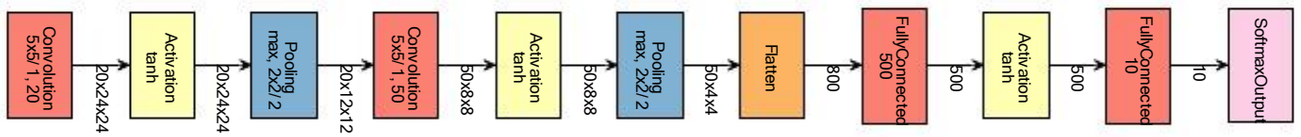
http://josephpcohen.com/w/visualizing-cnn-architectures-side-by-side-with-mxnet/

Share load across edge and cloud by DNN partitioning

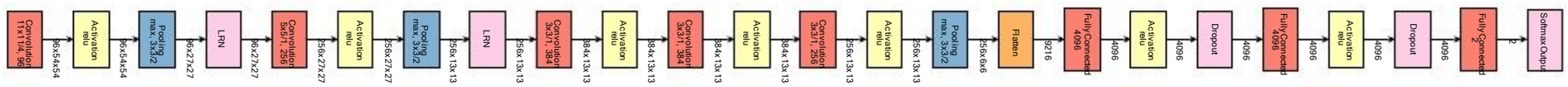


How do we decide the slicing point?

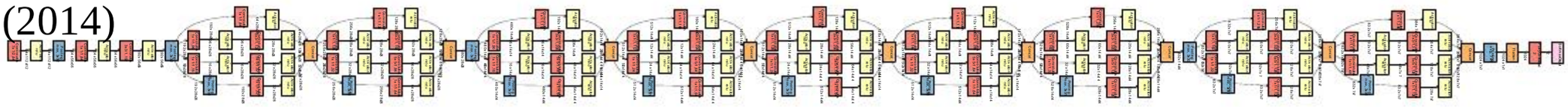
LeNet (1998)



AlexNet (2012)



GoogLeNet (2014)

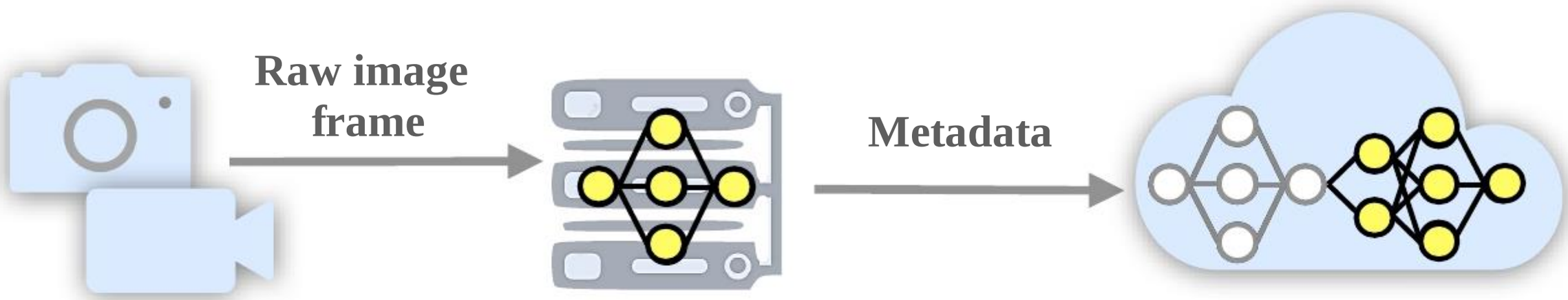


Inception V3 (2015)

ResNet (2015)

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

Share load across edge and cloud by DNN partitioning

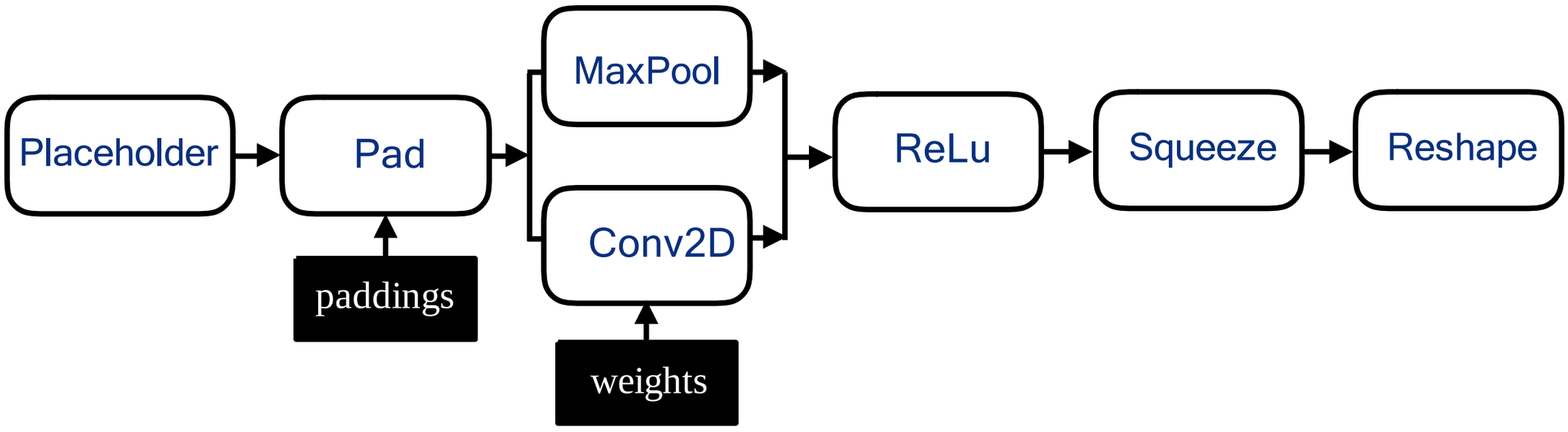


How do we decide the partition point?

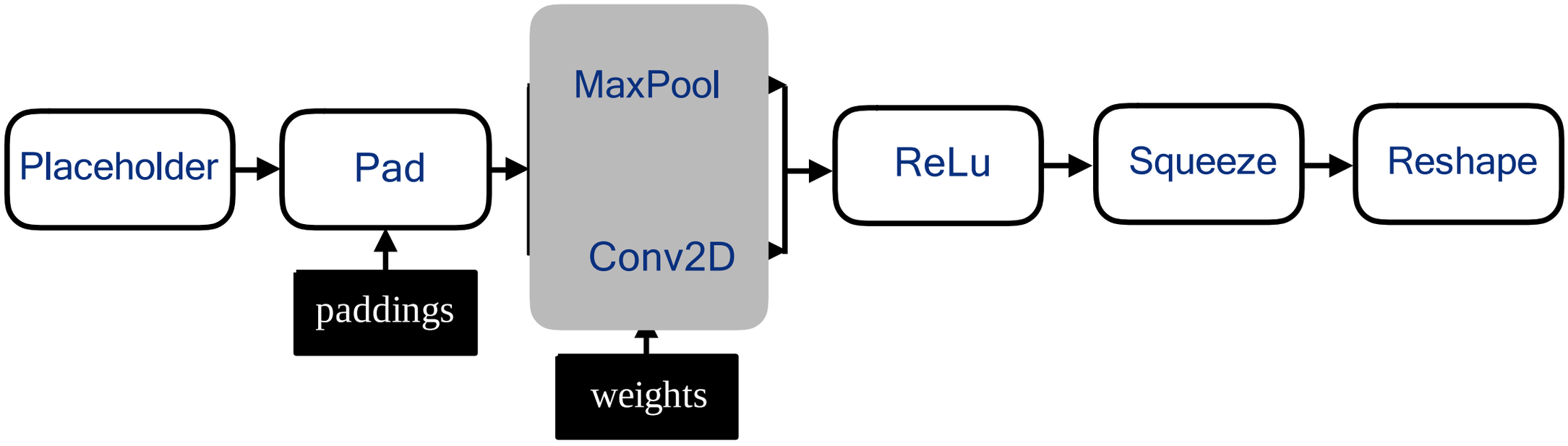
1. Filter out splittable candidates in DNN

2. Pick up a right one among the candidates

Listing splicing candidates

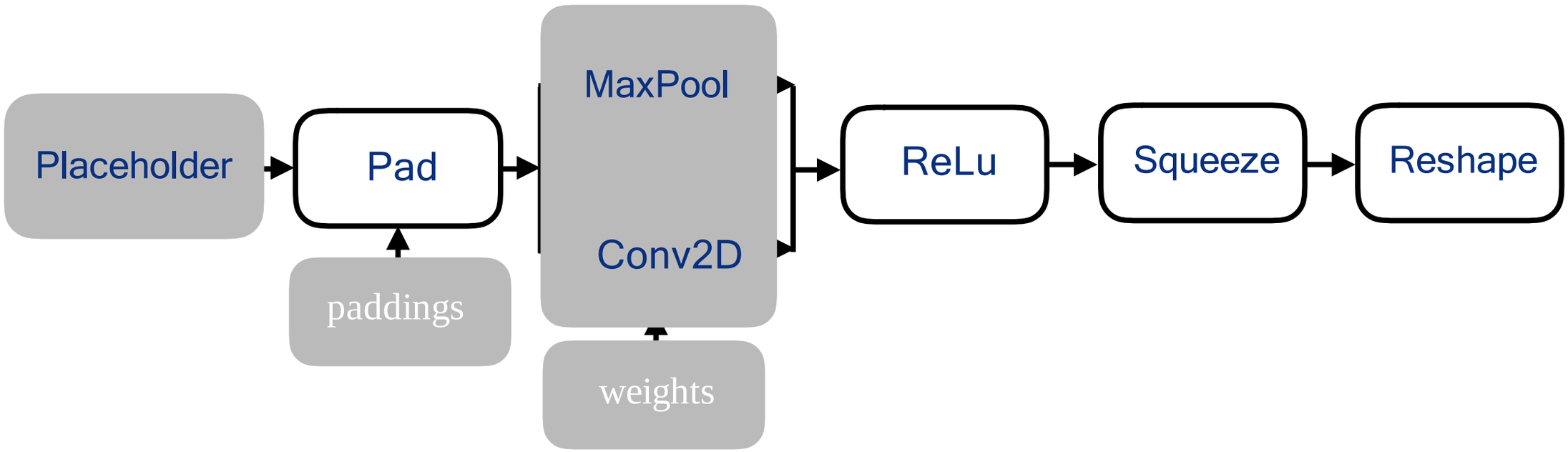


Listing splicing candidates



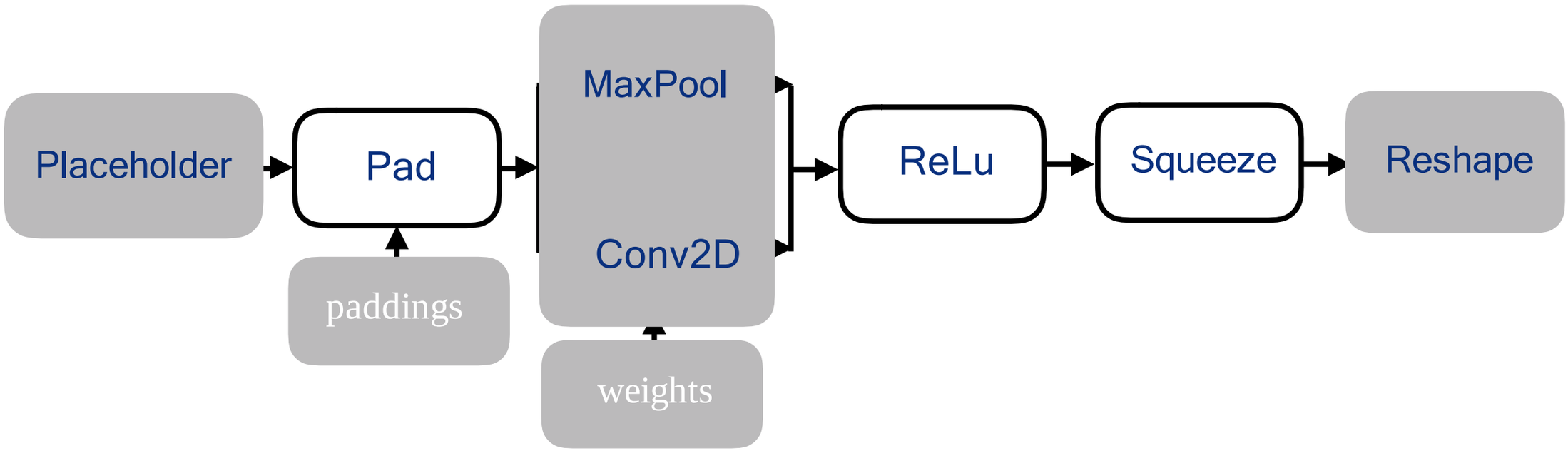
✗ Multi-parallel path

Listing splicing candidates



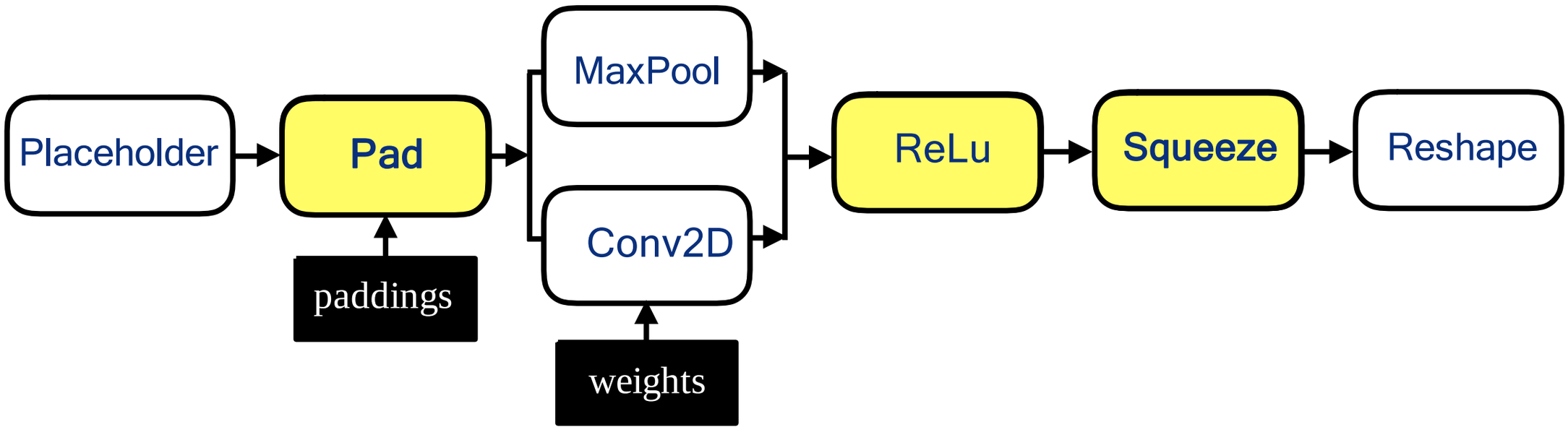
- ✗ Multi-parallel path
- ✗ Constant or reading operator

Listing splicing candidates



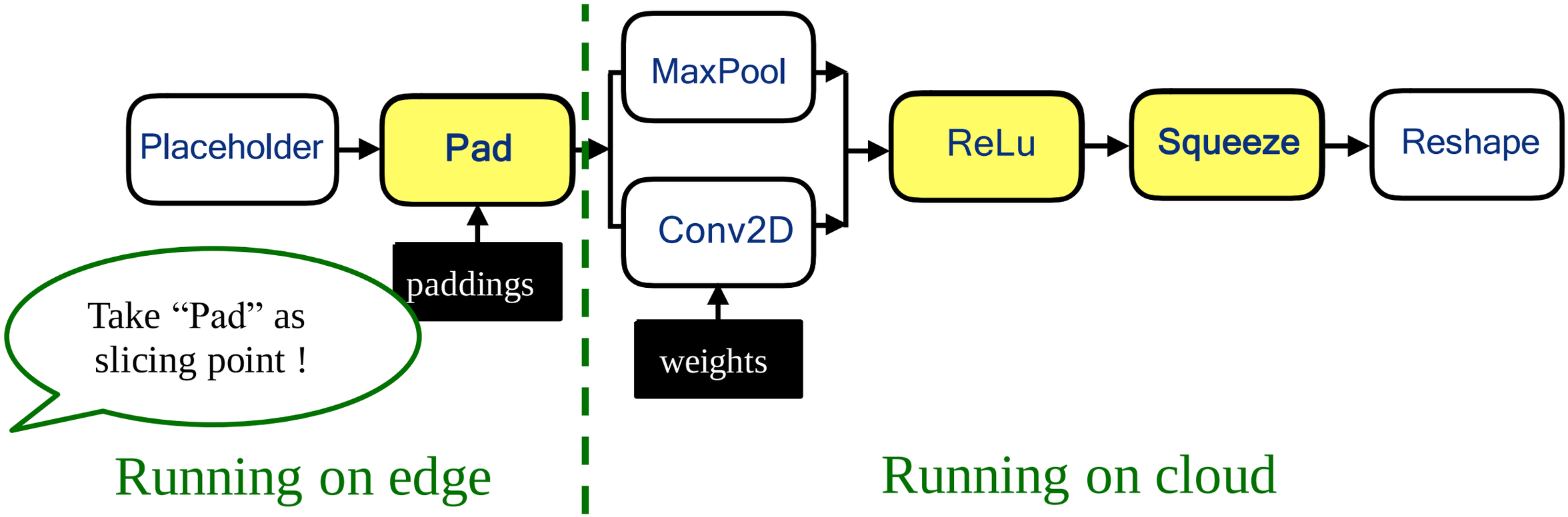
- ✗ Multi-parallel path
- ✗ Constant or reading operator
- ✗ Last operator

Listing splicing candidates



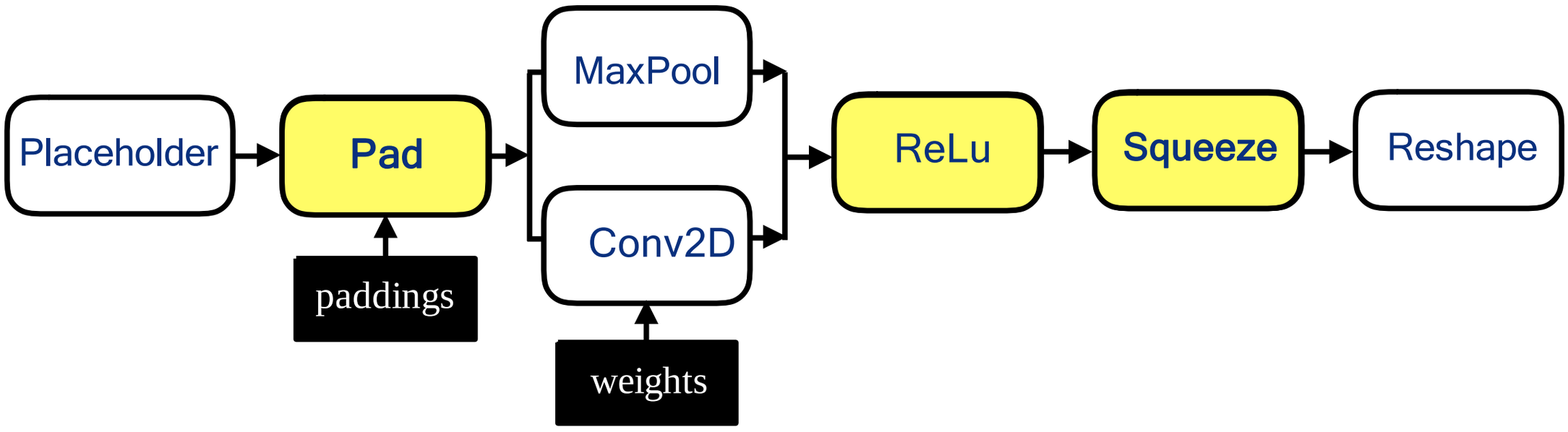
- ✗ Multi-parallel path
- ✗ Constant or reading operator
- ✗ Last operator

Listing splicing candidates

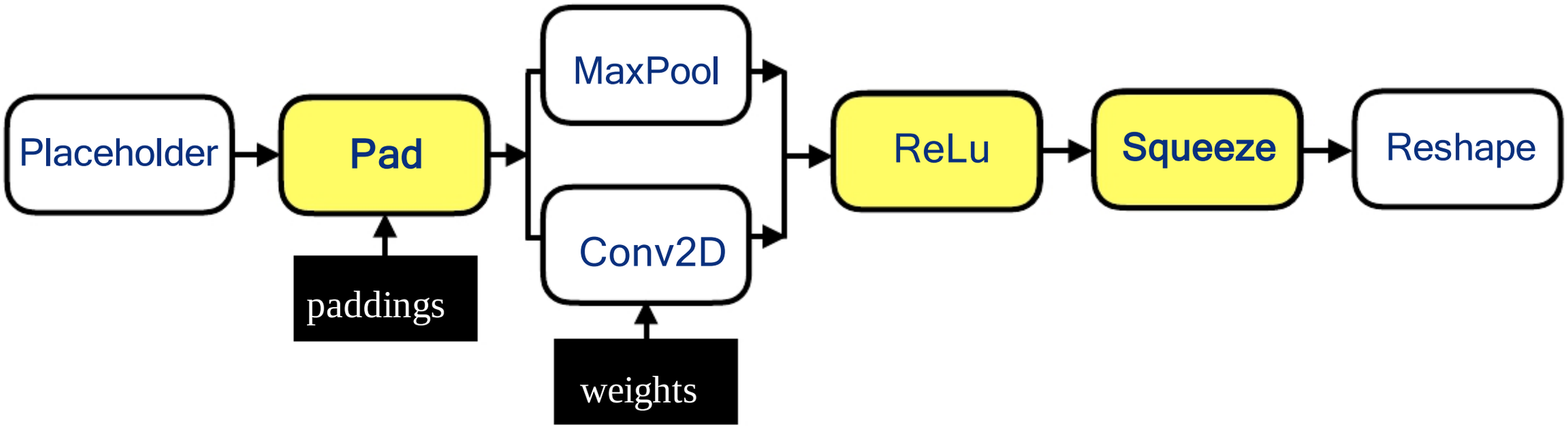


- ✗ Multi-parallel path
- ✗ Constant or reading operator
- ✗ Last operator

Evaluating splicing candidates

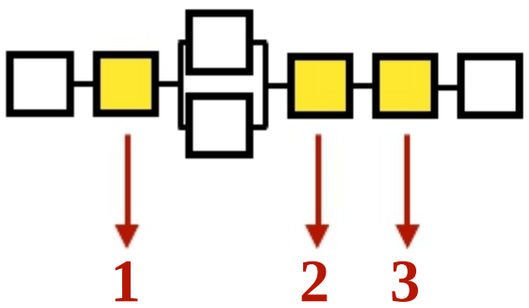


Evaluating splicing candidates

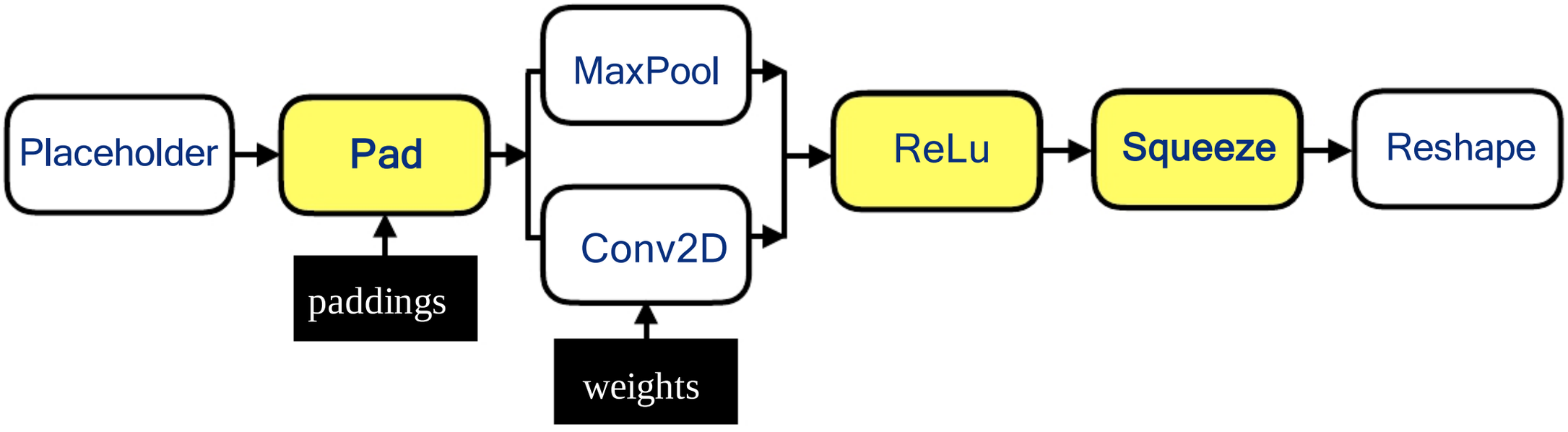


Strongman

Evaluate
every candidate

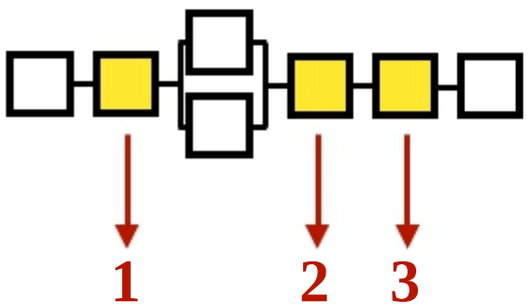


Evaluating splicing candidates



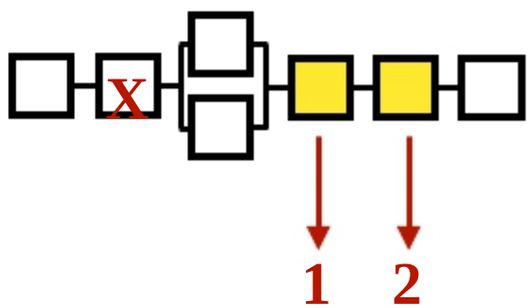
Strongman

Evaluate every candidate

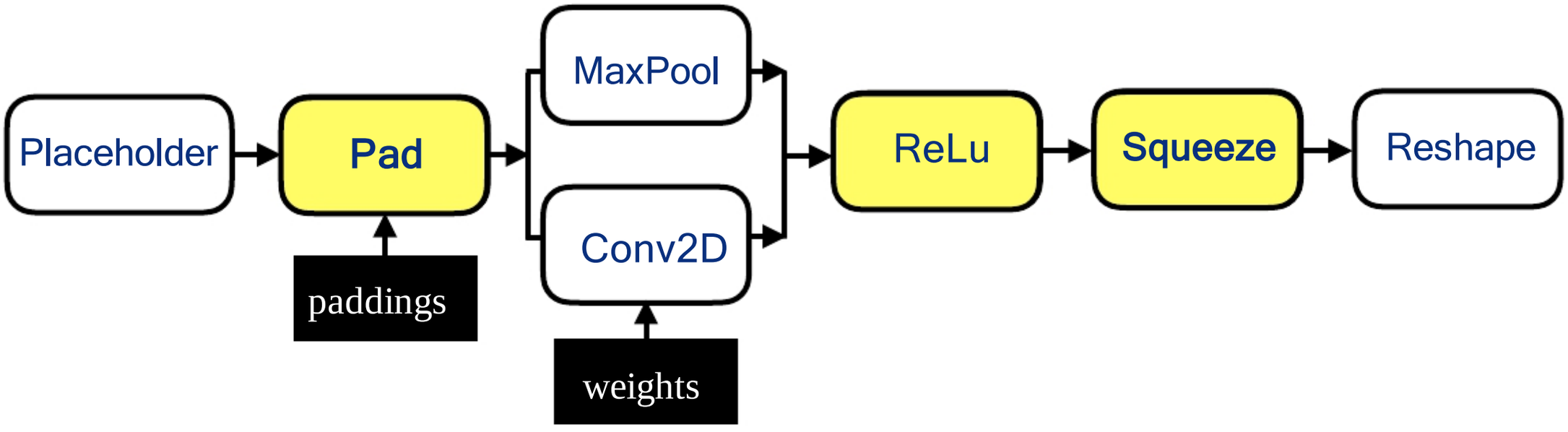


Comm-slim

Bypass candidates with high networking cost

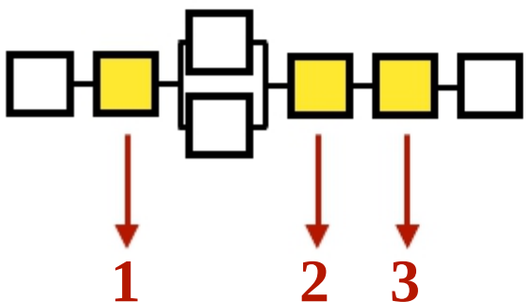


Evaluating splicing candidates



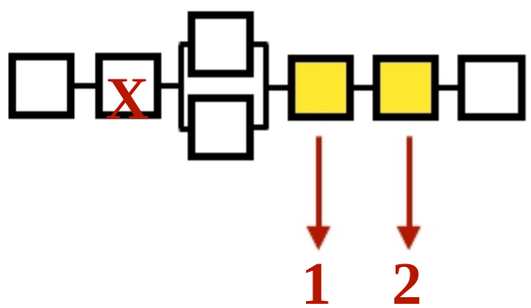
Strongman

Evaluate every candidate



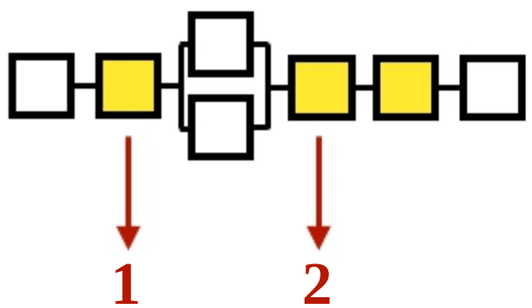
Comm-slim

Bypass candidates with high networking cost

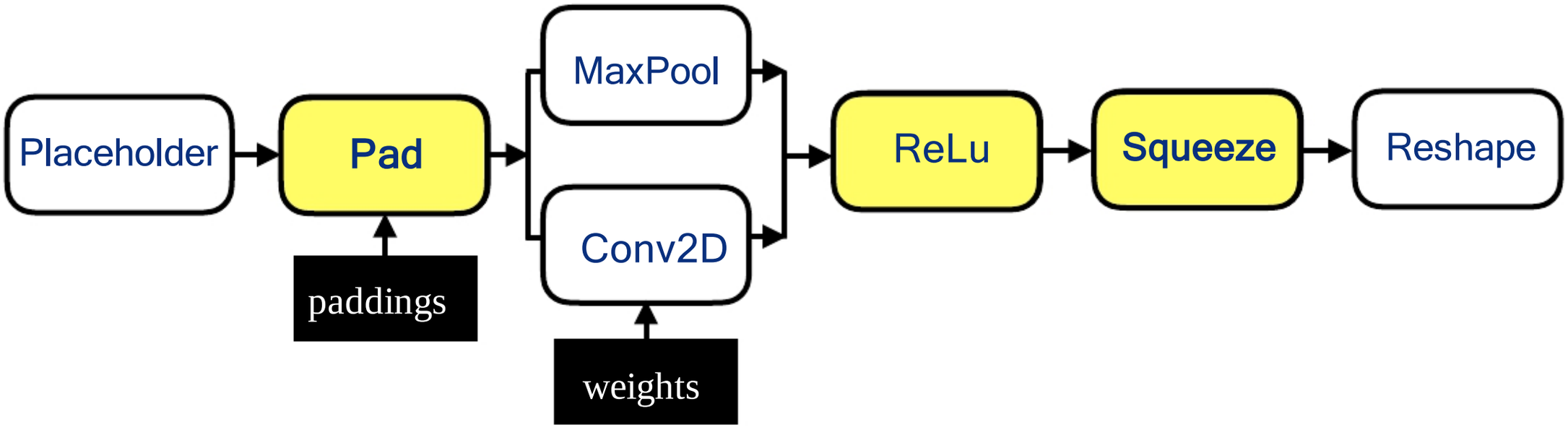


Early-stop

Stop evaluation when edge is overload

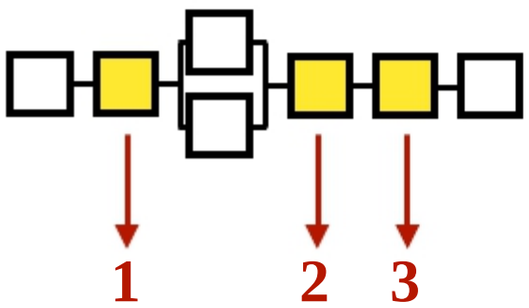


Evaluating splicing candidates



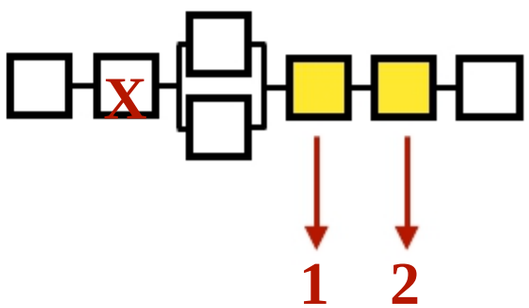
Strongman

Evaluate every candidate



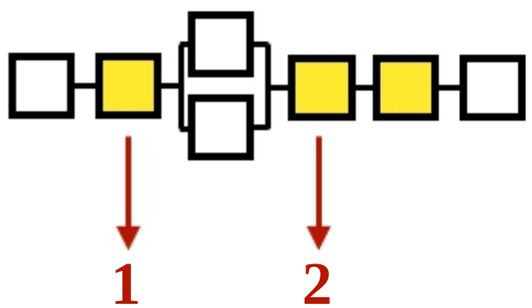
Comm-slim

Bypass candidates with high networking cost



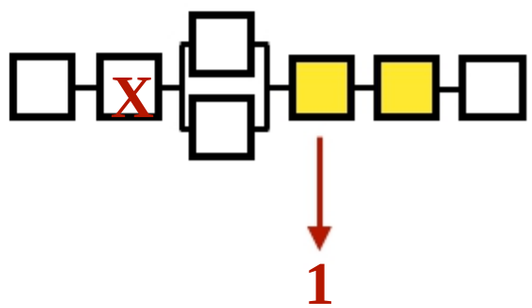
Early-stop

Stop evaluation when edge is overload

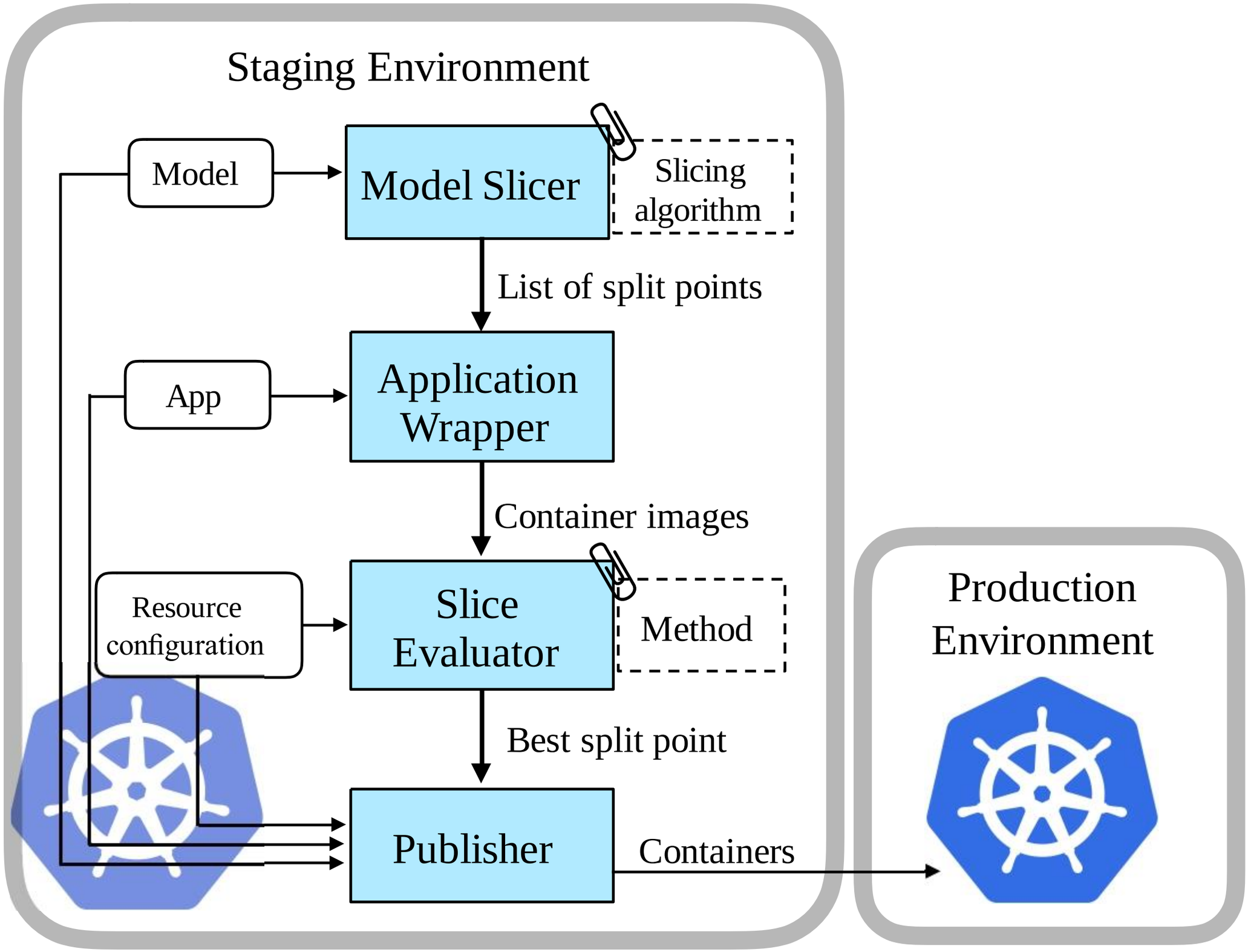


Hybrid

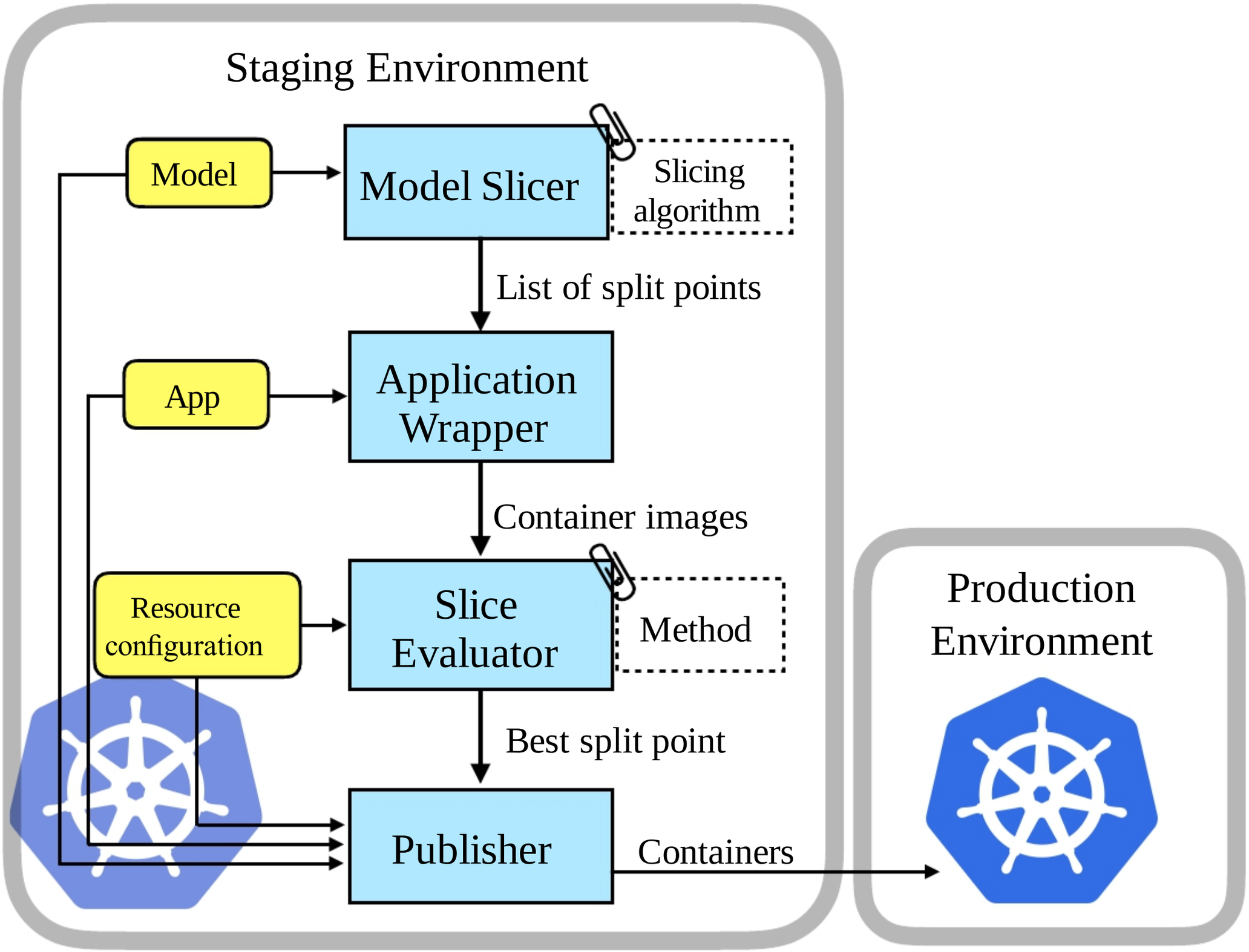
Combination of comm-slim and early-stop



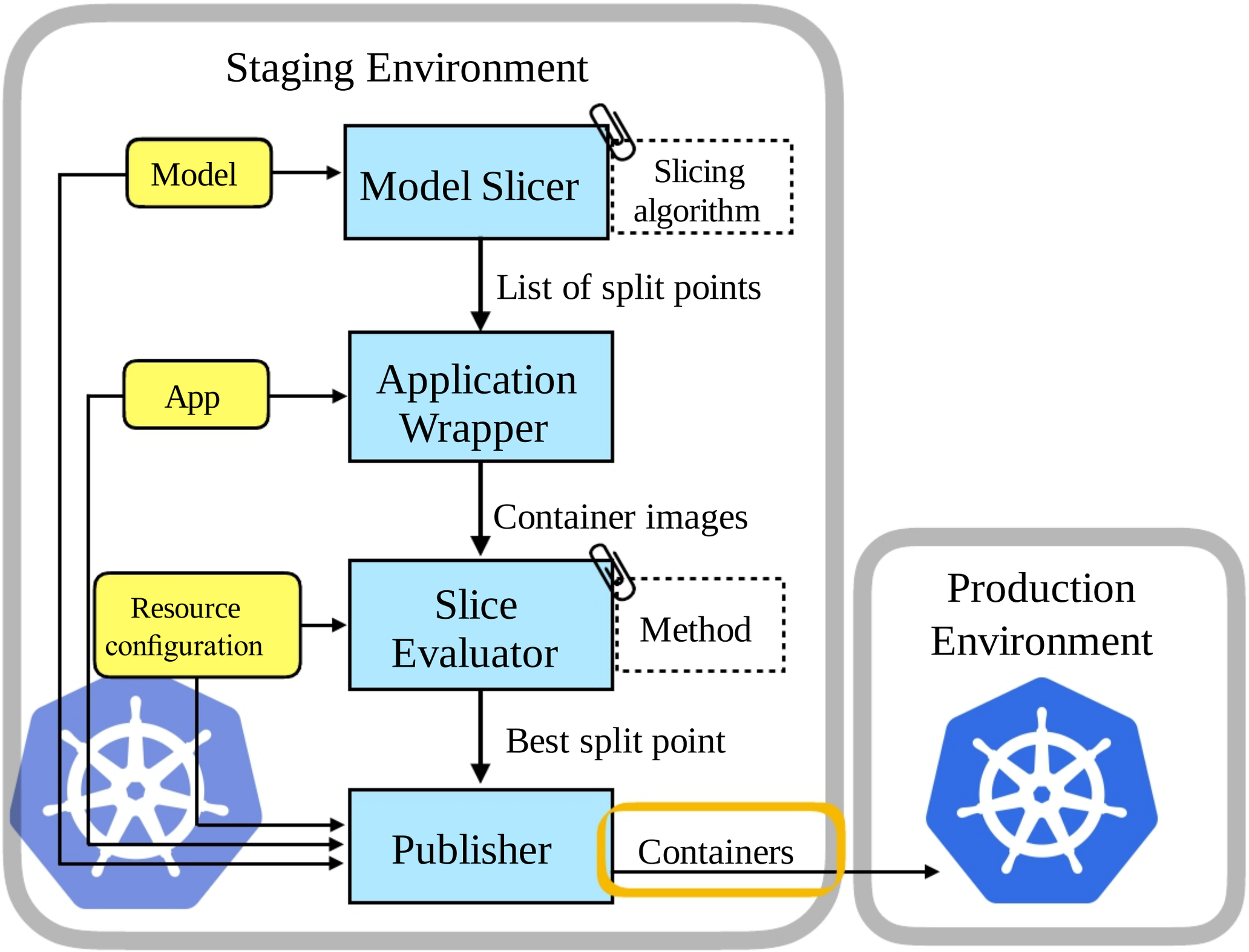
Couper Overview



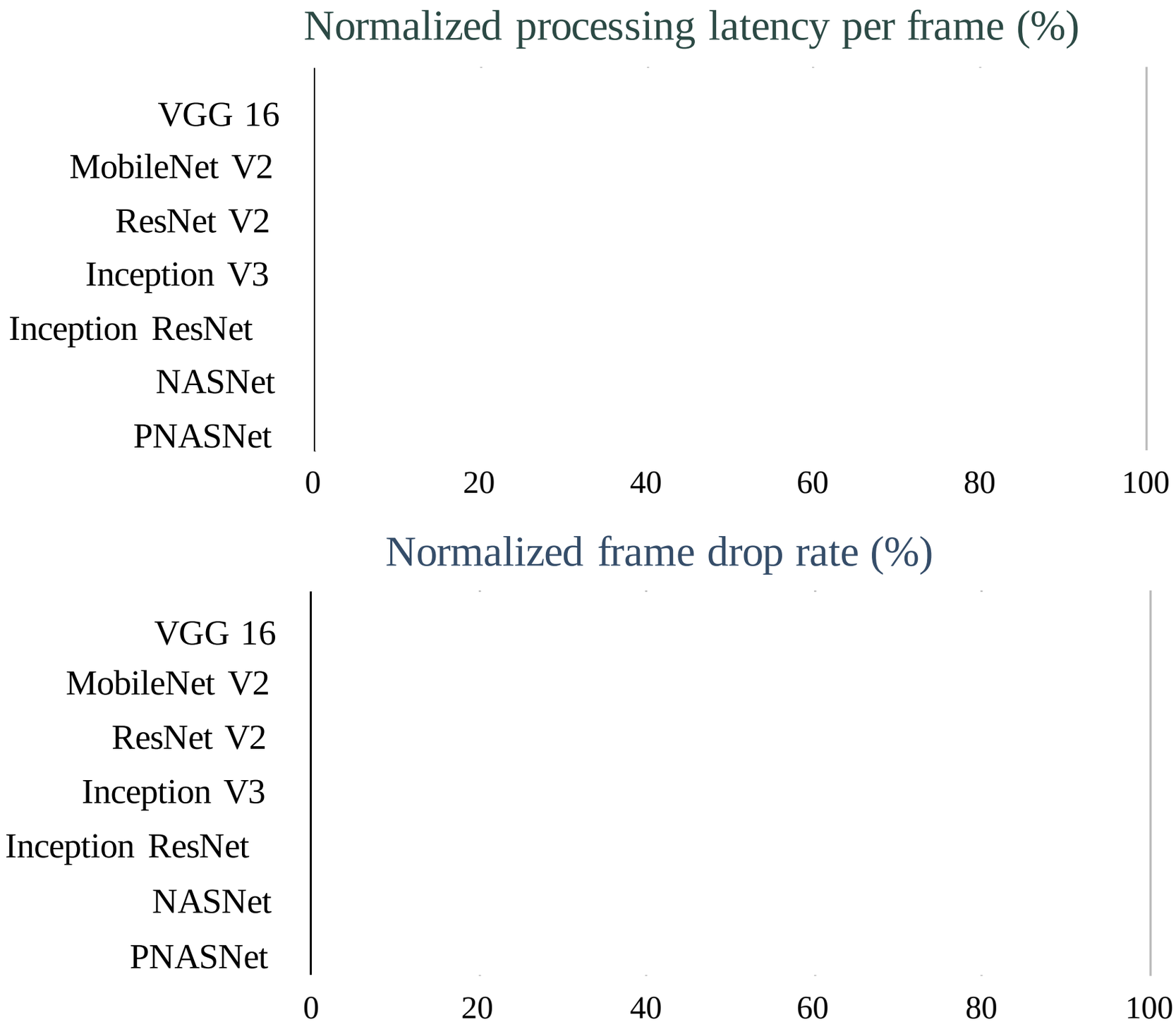
Couper Overview



Couper Overview

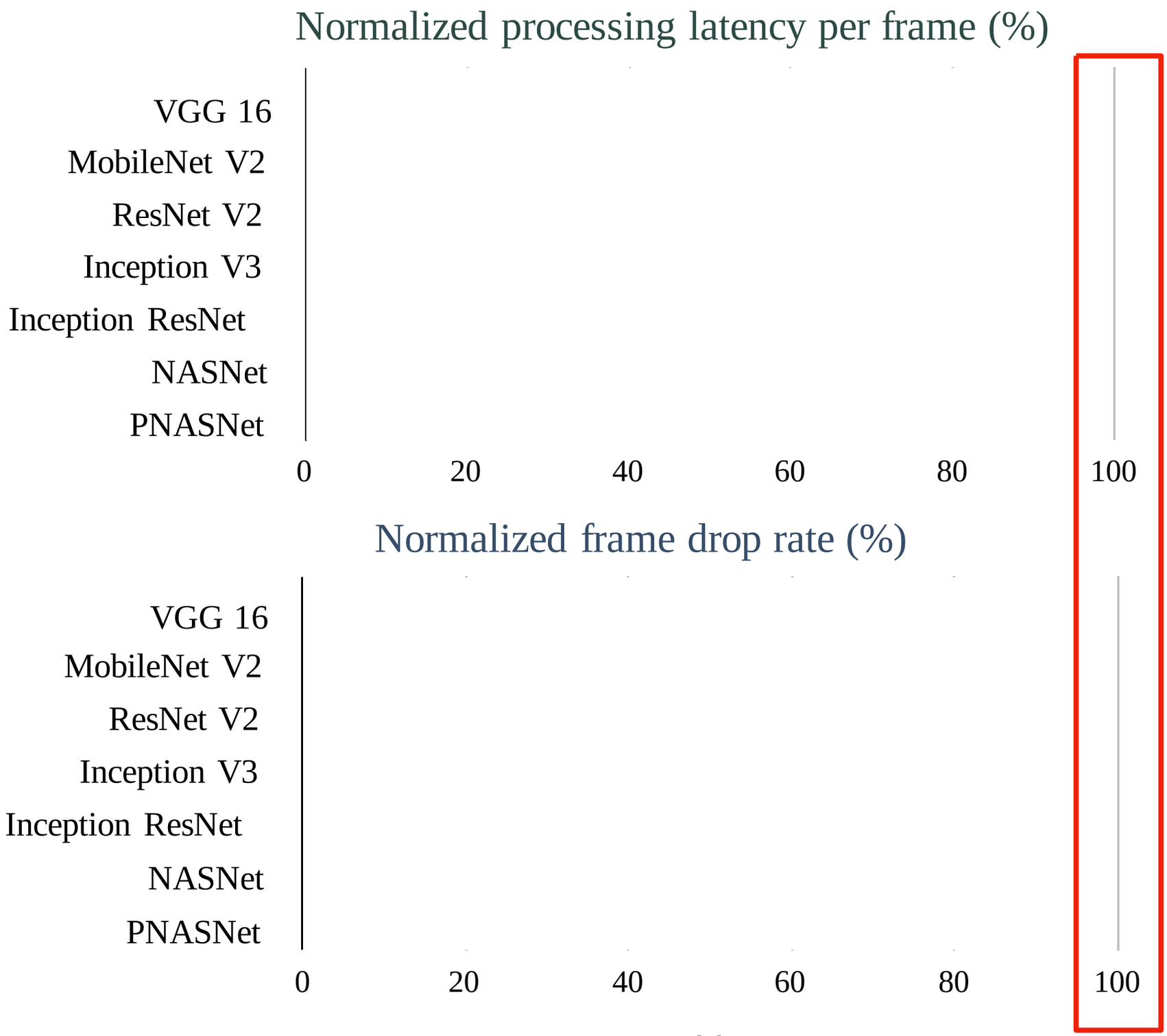


Results for different SLAs

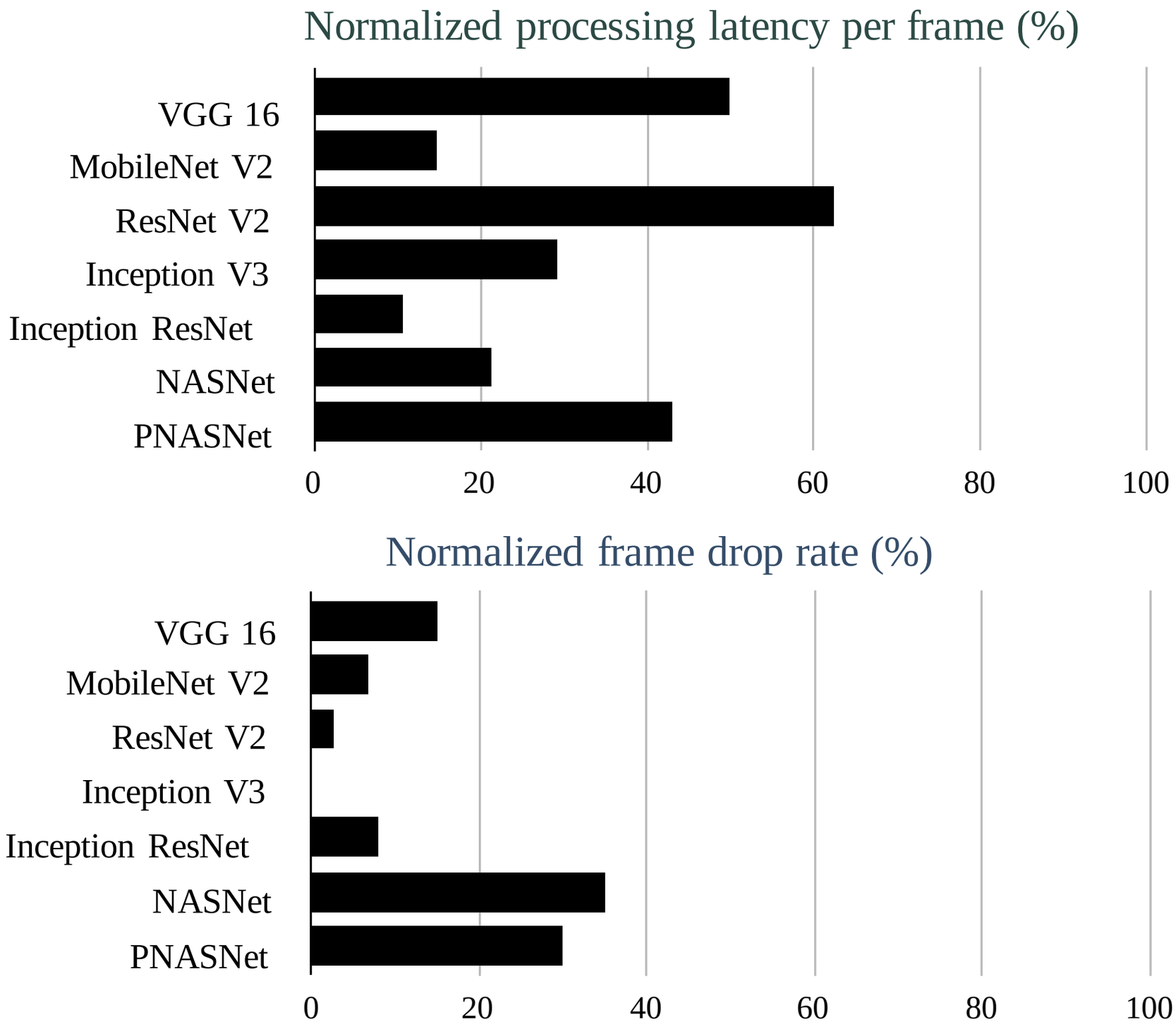


Results for different SLAs

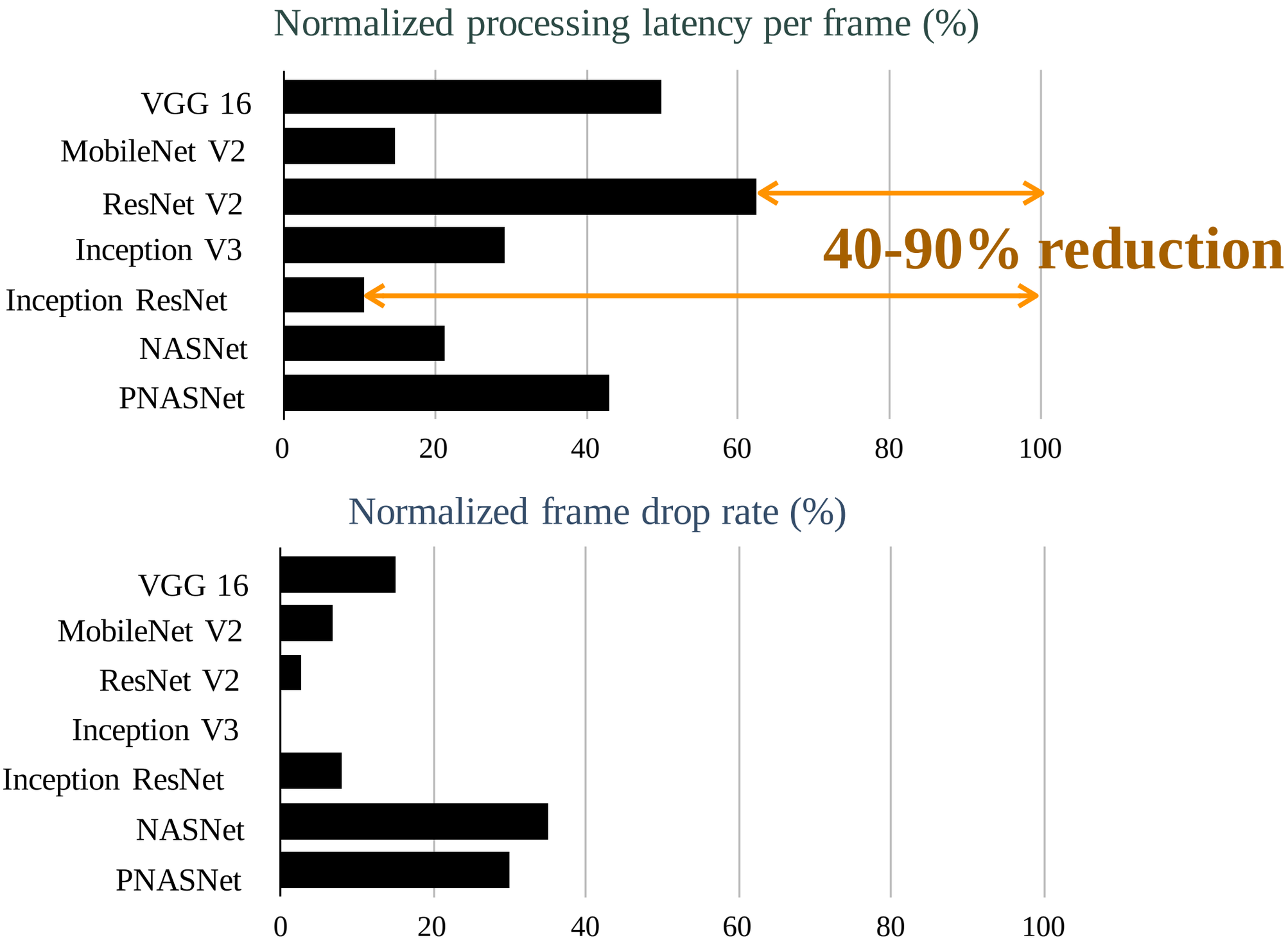
placing all DNN inference on cloud



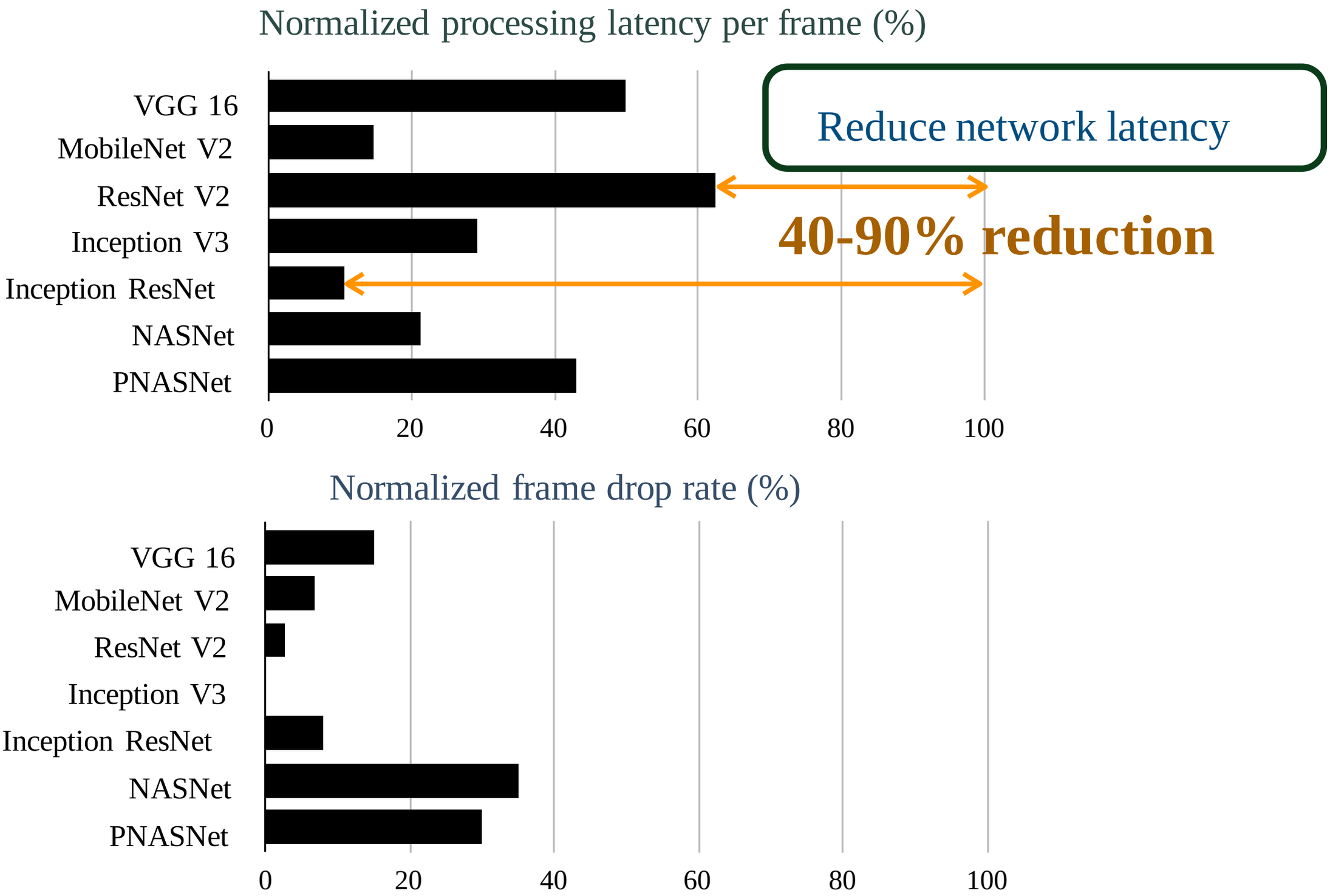
Results for different SLAs



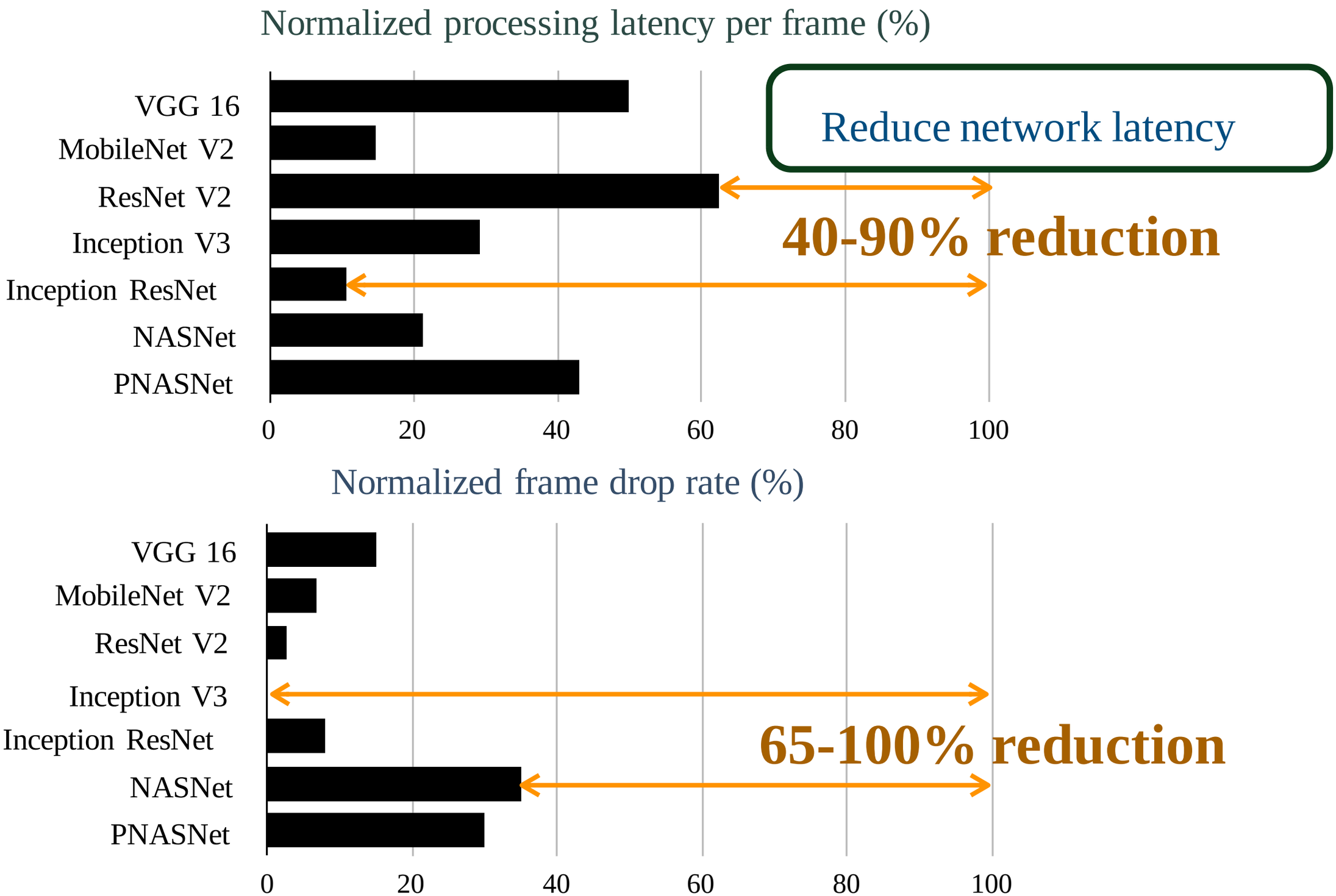
Results for different SLAs



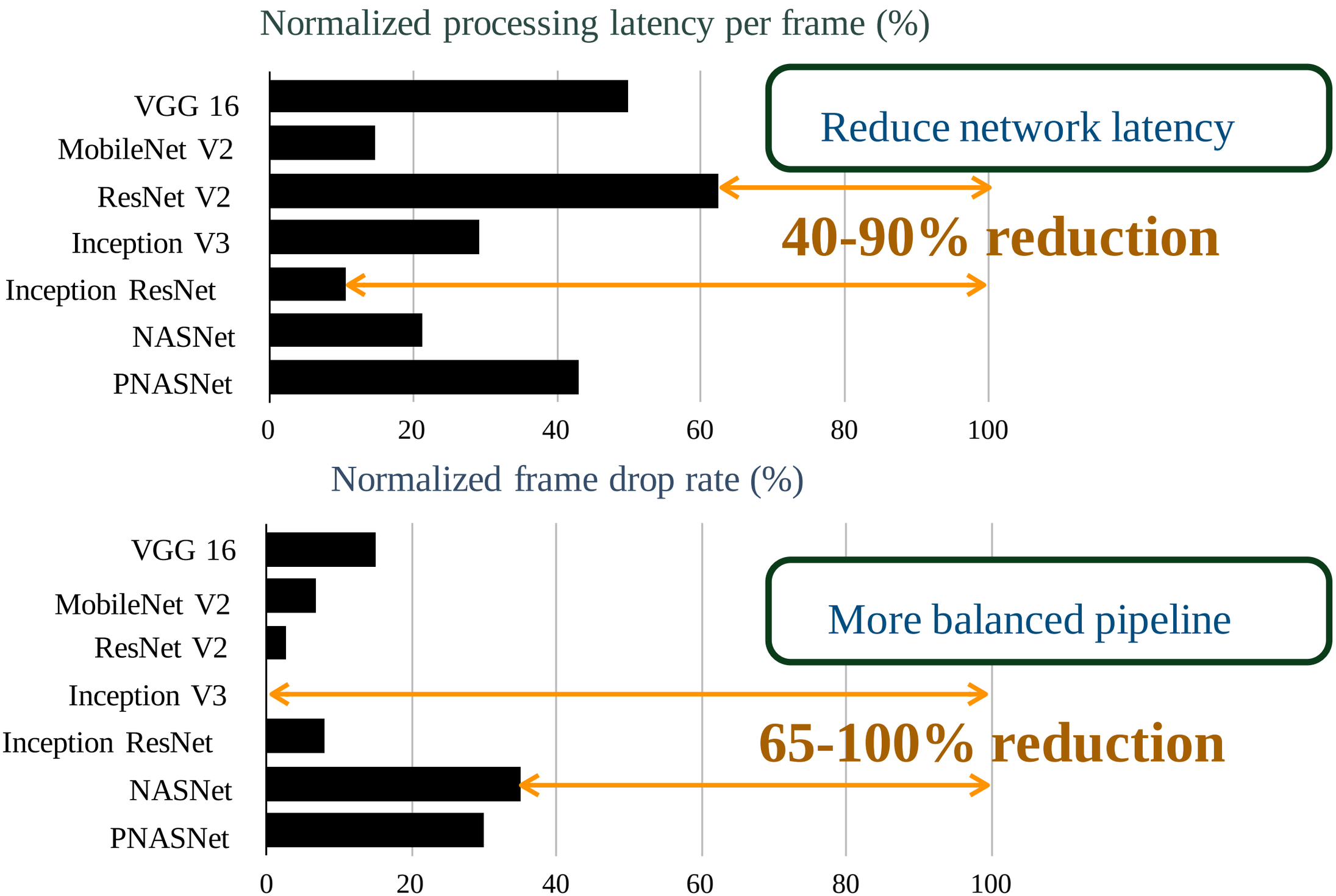
Results for different SLAs



Results for different SLAs



Results for different SLAs



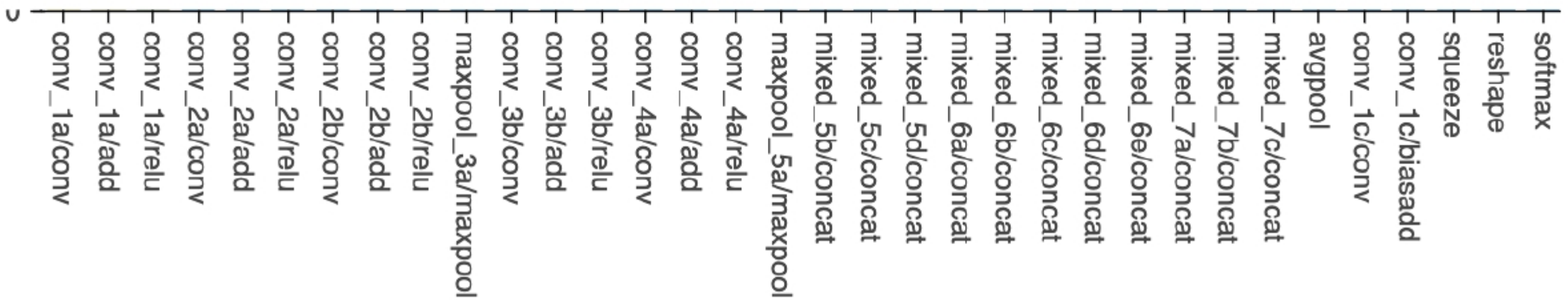
Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2

Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2

99% reduction

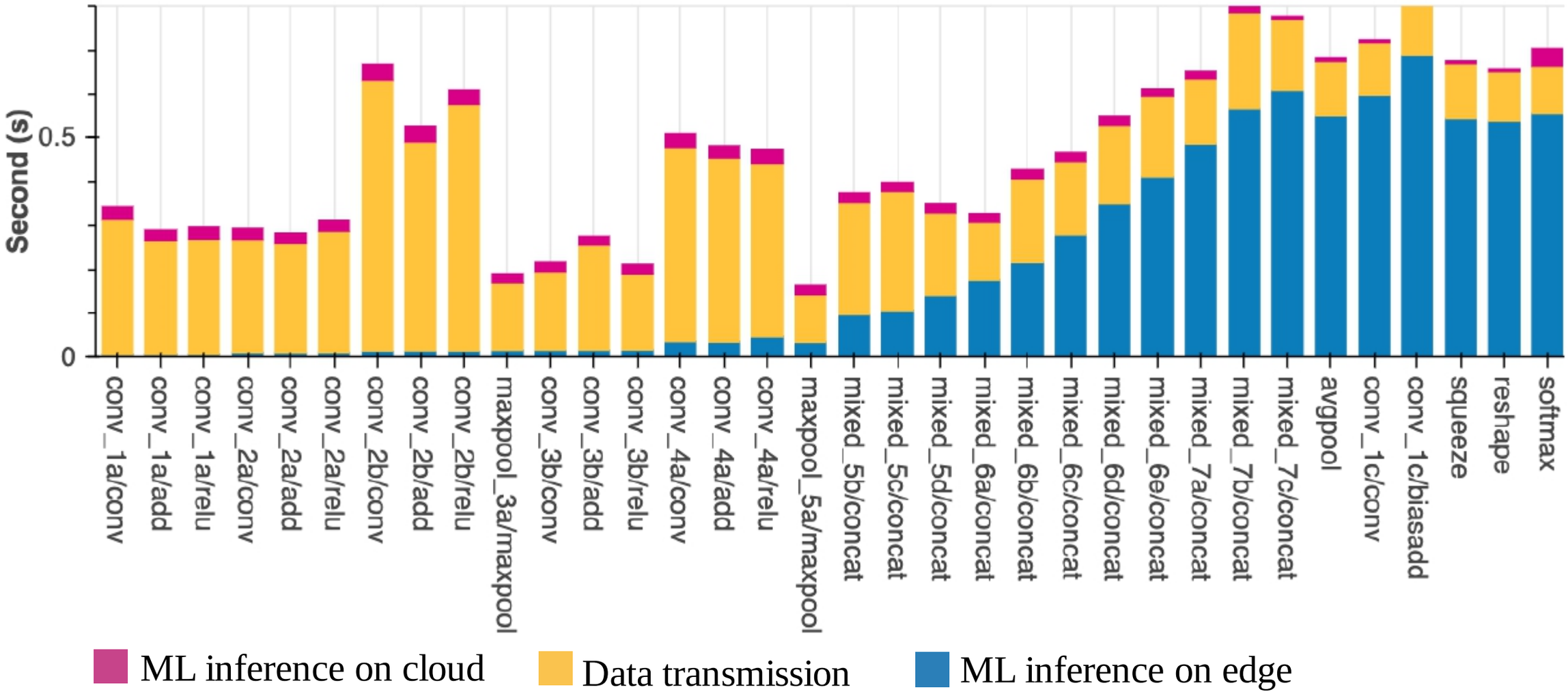
Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2

99% reduction

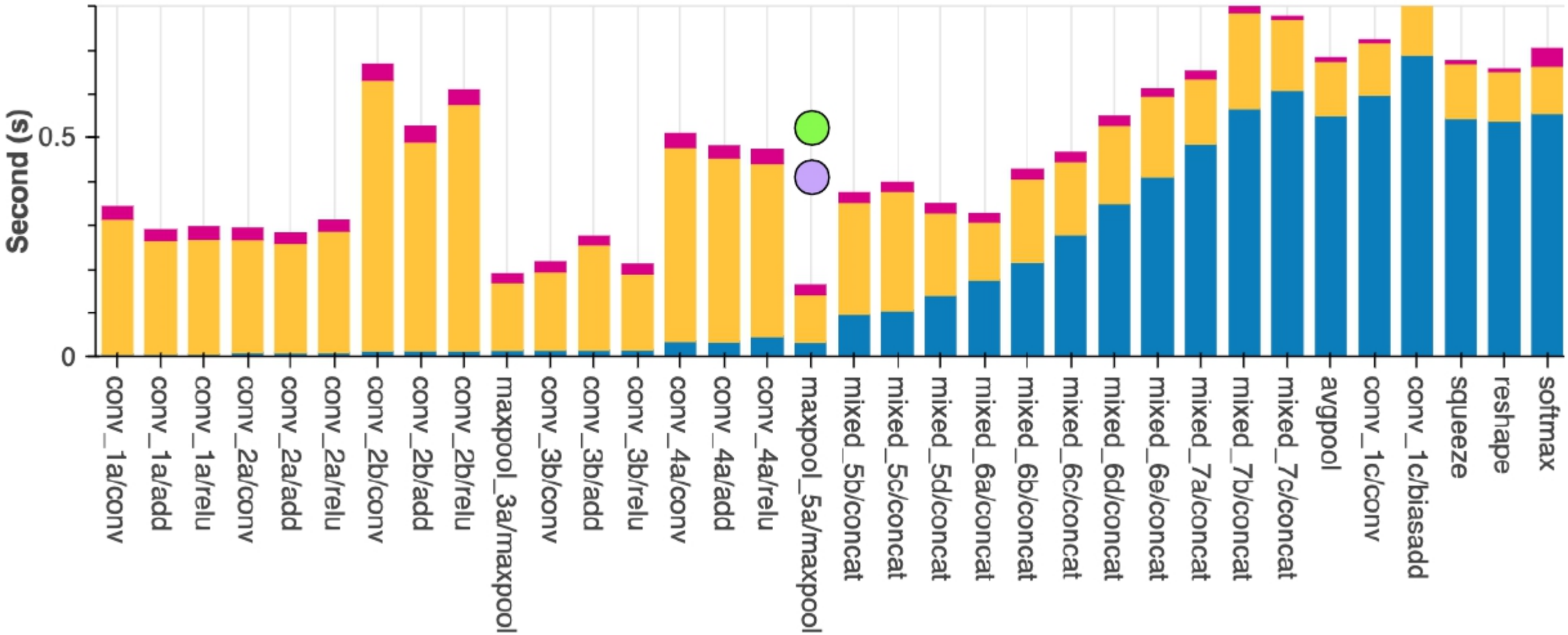


Strongman method tests 34 slicing candidates

Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2

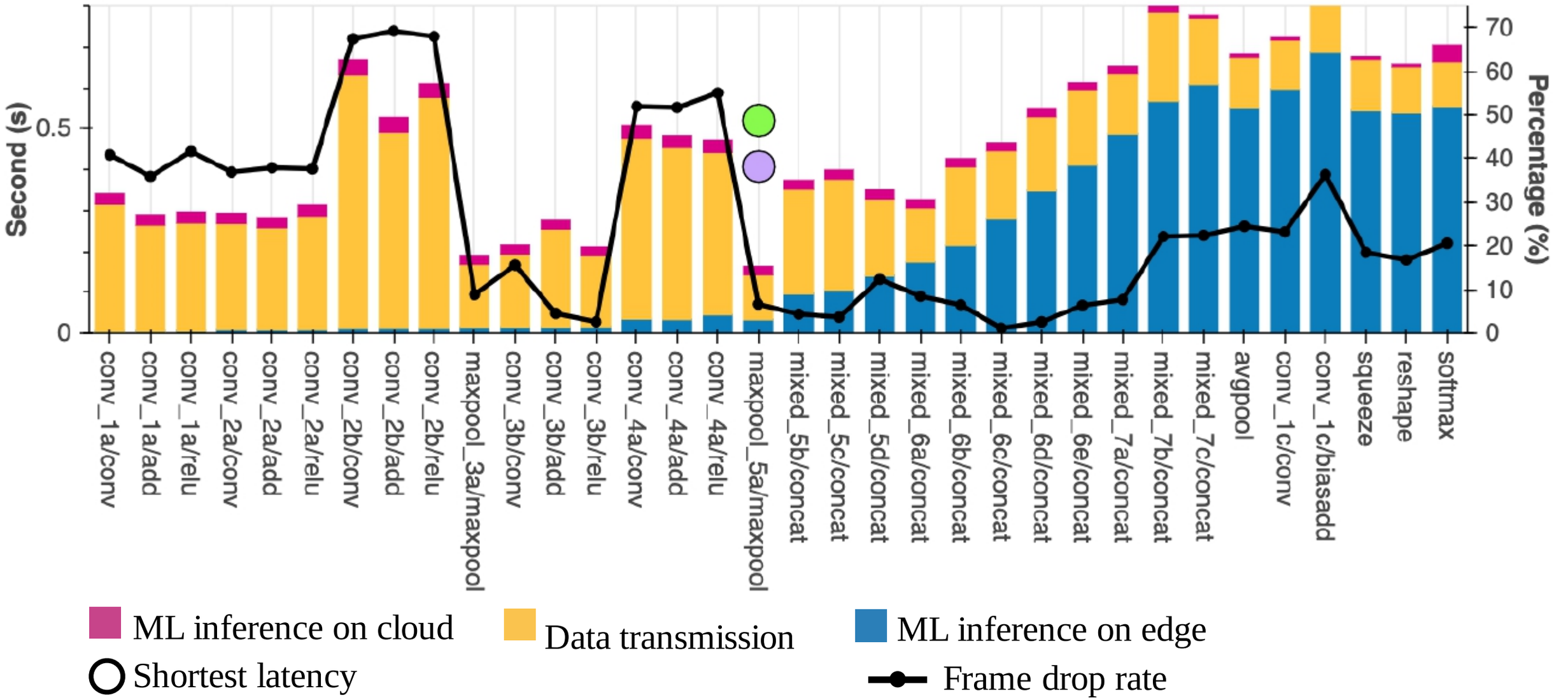


Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2

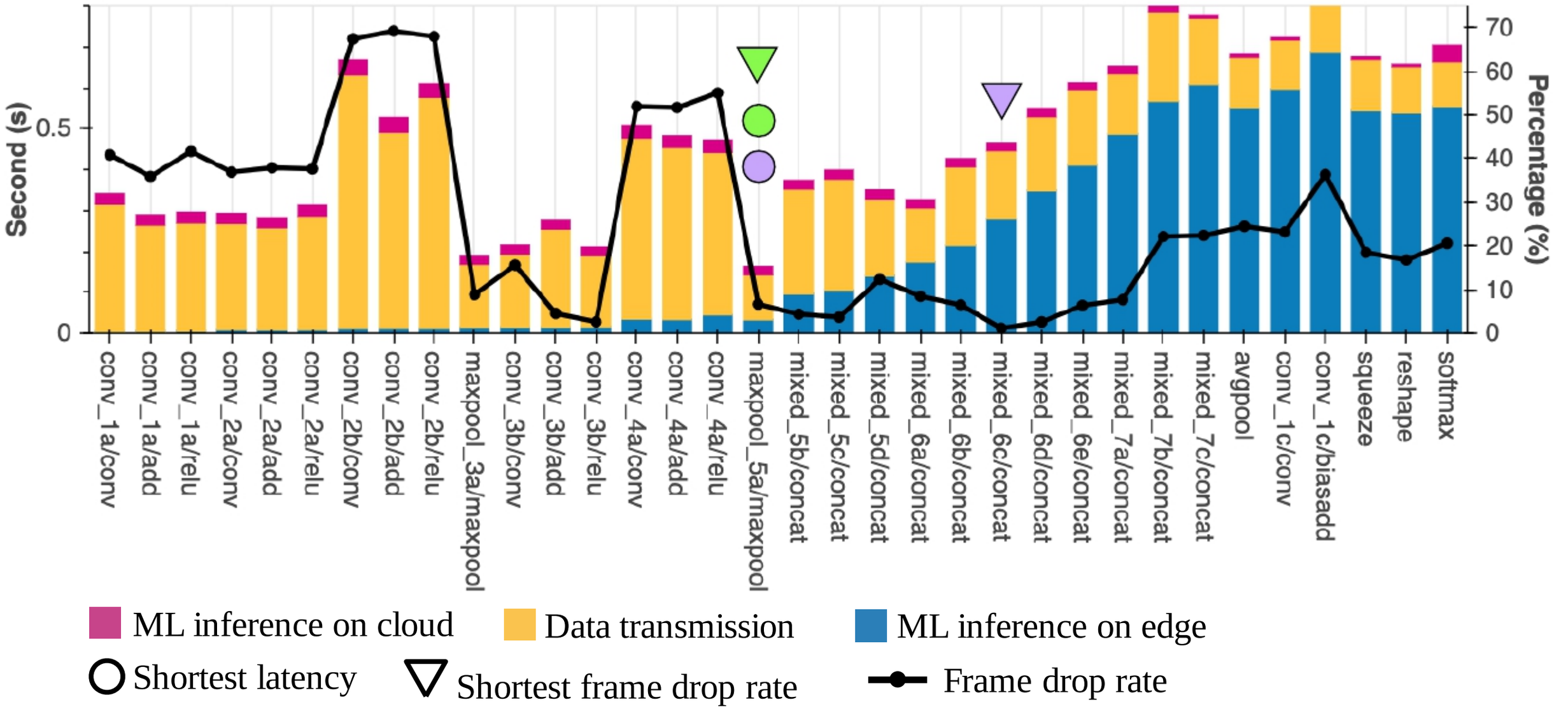


Hybrid method can find the same slicing deployment with much smaller problem space

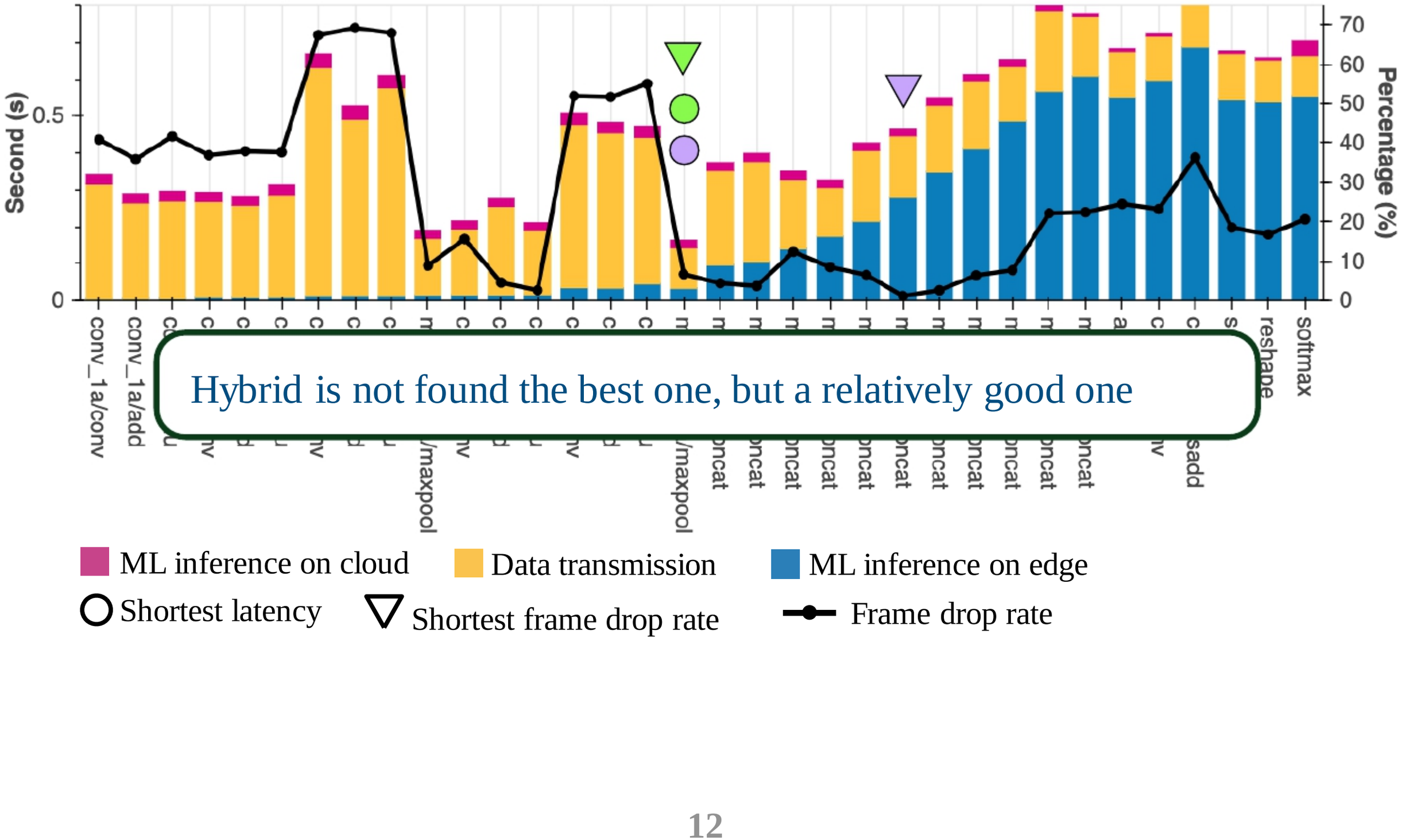
Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2



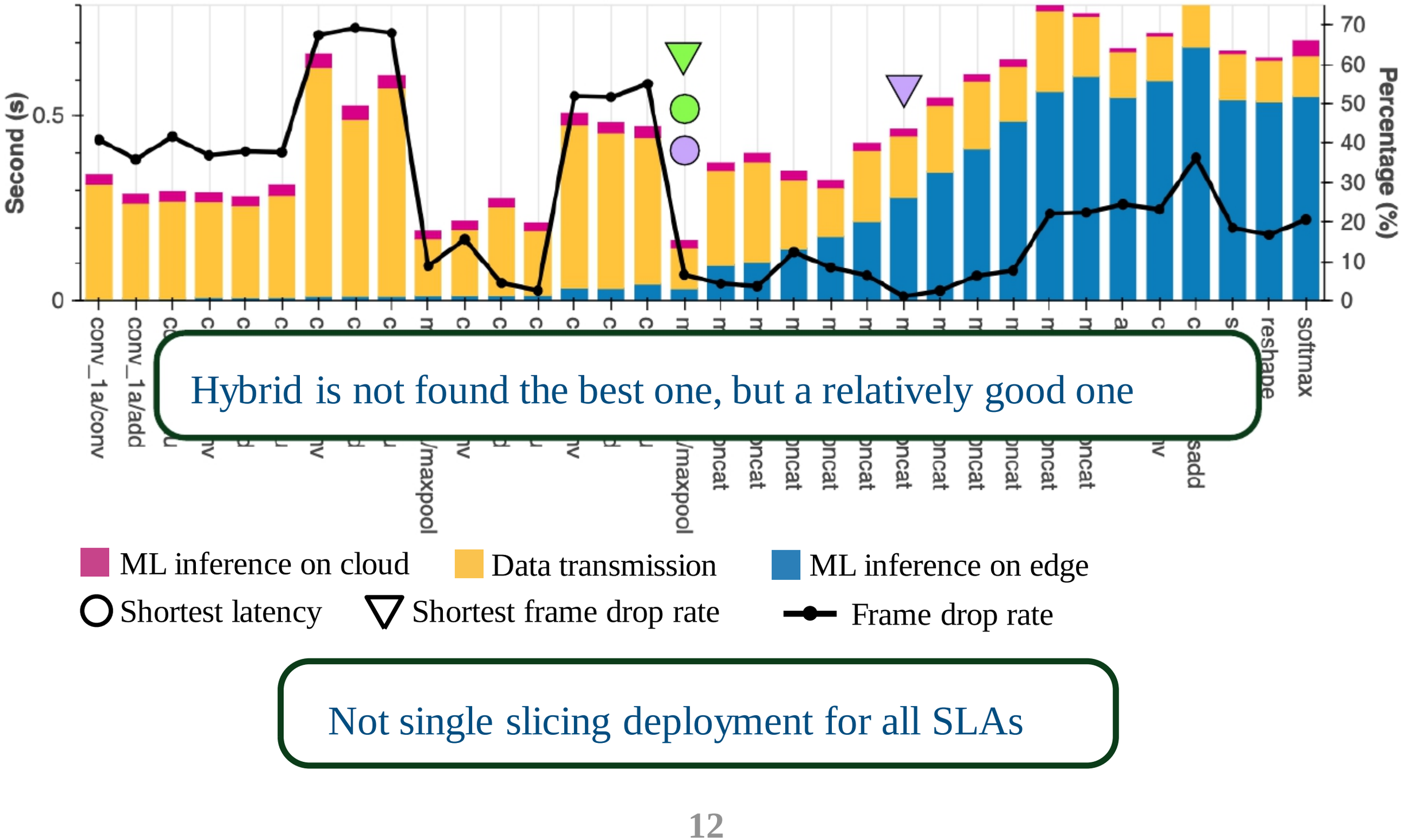
Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2



Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2



Model	#Operator	Method	
		Strongman	Hybrid
InceptionV3	788	34	2



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.

Other Opinions

- 文章提供了一个新思路，从DNN网络模型的共性特征出发，寻找普适的分割算法，适应模型层出不穷的时代
- 测试的模型数量比较多，实验数据丰富

- 文章的架构是单向的端->边->云，而实际还有很多应用是端->边->云->边->端，不能完全照搬
- 实验环节中的终端到云端的RTT只有50-60ms，不能匹配所有的应用场景

- 项目开源，我们可以在项目场景下借用这个工具进行多个模型的分割点寻找和对比，或者是基于该项目设计更加完善的自动分割算法

Thanks for your attention!